UNIVERSIDADE FEDERAL DO PARANÁ

RODRIGO RENE MENEGAZZO

MONTAGEM, ANOTAÇÃO E COMPARAÇÃO DO GENOMA PARCIAL DA BACTÉRIA *Azoarcus olearius* DQS-4^{T}

CURITIBA

2015

RODRIGO RENE MENEGAZZO

MONTAGEM, ANOTAÇÃO E COMPARAÇÃO DO GENOMA PARCIAL DA BACTÉRIA *Azoarcus olearius DQS4*

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, como requisito parcial para obtenção do título de Mestre em Bioinformática

Orientador: Dr. Helisson Faoro

Co-orientador: Prof. Dr. Emanuel Maltempi de Souza

CURITIBA

2015

Menegazzo, Rodrigo Rene
M541 Montagem, anotação e comparação do genoma parcial da bactéria *Azoarcus*
olearius DQS4 / Rodrigo Rene Menegazzo. - Curitiba, 2015.
88 f.: il., tabs, grafs.

Orientador: Helisson Faoro
Co-orientador: Emanuel Maltempi de Souza
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. *Azoarcus olearius. 2*. Nitrogênio - Fixação. 3. Genética - Processamento de dados. 4. Bioinformática. I. Faoro, Helisson. II. Souza, Emanuel Maltempi de. III. Título. IV. Universidade Federal do Paraná.

CDD 575.113

TERMO DE APROVAÇÃO

RODRIGO RENE MENEGAZZO

"Montagem, anotação e comparação do genoma parcial da bactéria *Azoarcus olearius* DQS-4^T"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Dr. Helisson Faoro

Coorientador: Prof. Dr. Emanuel Malpempi de Souza

Dr. Leonardo Magalhães Cruz
Universidade Federal do Paraná - UFPR

Dr. Vinicius Almir Weiss
Bolsista PNPD/CAPES/Projeto Biologia Computacional - UFPR

Curitiba, 25 de março de 2015

AGRADECIMENTOS

Todos que estiveram de alguma forma envolvidos na realização deste trabalho;

Em especial aos meus pais, colegas e meus orientadores, Helisson Faoro e Emanuel Maltempi de Souza.

CNPq, INCT para Fixação Biológica de Nitrogênio;

Coordenação do curso de Pós-graduação em Bioinformática**.**

RESUMO

Existem na natureza bactérias capazes de transformar o nitrogênio existente no ar em amônia, um processo denominado Fixação Biológica de Nitrogênio (FBN), e transferi-lá para plantas nessa forma a qual consigam utilizar. Esta associação entre plantas e bactérias é conhecida desde o início do século XX para plantas leguminosas, na qual as plantas infectadas por bactérias capazes de realizar a fixação biológica de nitrogênio desenvolvem um nódulo radicular onde o processo ocorre. Um dos maiores desafios destes últimos anos foi encontrar bactérias que realizassem fixação biológica de nitrogênio com plantas não leguminosas, principalmente gramíneas e cereais, nas quais não há formação de nódulo. Bactérias isoladas do gênero *Azoarcus* possuem a capacidade de realizar esta associação com plantas além de promoverem o crescimento vegetal. Em 2013, foi publicada a descoberta de uma nova bactéria do gênero Azoarcus, denominada *Azoarcus olearius* estirpe DQS4. Esta publicação despertou o interesse no estudo da biologia desta bactéria. Para tanto foi realizado o sequenciamento de seu genoma e posterior montagem, anotação e comparação com as demais bactérias do gênero *Azoarcus*. O genoma do *A. olearius* DQS4 é formado por um cromossomo circular de 4,45 Mb distribuídos em 4.067 genes codificados. O cálculo de sua identidade média nucleotídea mostrou 99% de identidade com o *Azoarcus* sp. BH72, 82,64% com o *Azoarcus* sp. KH32, 82,45% com o *Azoarcus* sp. EbN1 e 82,79% com *Azoarcus toluclasticus*. As análises de sintenía genômica também mostraram uma alta relação com o genoma de *Azoarcus* sp. BH72. Os genes necessários para a fixação de nitrogênio foram encontrados no genoma de *A. olearius* DQS4 e apresentaram alta similaridade com genes envolvidos na fixação de nitrogênio em *Azoarcus* sp. BH72. Esta alta similariedade entre o *A. olearius* DQS4 e o *Azoarcus* sp. BH72 sugere que estas duas bactérias podem ser categorizadas como diferentes estirpes dentro da mesma espécie.

Palavras-chave: *Azoarcus olearius* DQS4, Sequenciamento genômico, Fixação biológica de nitrogênio, comparação, *Azoarcus* sp. BH72

ABSTRACT

There are bacterias in nature capable of transforming the atmospheric nitrogen into ammonia, a process called Nitrogen Biological Fixation (NFB), allowing plants to use it. This association between plants and bacteria has been known since the early of twenty century for legumes, which plants infected by bacteria capable of performing biological nitrogen fixation developed a root nodule where the process occurs. One of the biggest challenges has been looking for the bacteria that perform biological nitrogen fixation with non-legumes, mainly grasses and cereal plants whitout nodulation. Bacteria from *Azoarcus* genus, that have the ability to perform this association with plants and to promote plant growth, where isolated. In 2013, the discovery of a new species of the *Azoarcus* genus, *Azoarcus olearius* strain DQS4 was published. Due to the importance of *Azoarcus* genus, the genome of *A. olearius* DQS4 was sequenced, assembled, annotated and compared with other bacteria of the *Azoarcus* genus. The *A. olearius* DQS4 has one circular chromosome of 4,45 Mb distributed in 4,067 encoded genes. Determination of the average nucleotide identity showed 99% of identity with *Azoarcus* sp. BH72, 82,64% with *Azoarcus* sp. KH32, 82,45% with *Azoarcus* sp. EbN1 and 82,79% to *Azoarcus toluclasticus*. Genome dotplot analysis also showed a high sinteny with *Azoarcus* sp. BH72. The genes required for nitrogen fixation were found in *A. olearius* DQS4 and are closely related to nitrogen-fixing genes of *Azoarcus* sp. BH72. This high similarity between *A. olearius* DQS4 and *Azoarcus* sp. BH72 suggested that these two bacteria might be categorized as different strains of the same species.

Key-words: *Azoarcus olearius* DQS4, Genome sequencing, Nitrogen Biological Fixation, comparison, *Azoarcus* sp. BH72.

LISTA DE SIGLAS

BLAST – Basic Local Alignment Search Tool

K-mer – (Semente) fração menor de uma sequência com tamanho pré-definido “k”

ORF – (Open Reading Frame) Fase de leitura aberta

pb – Pares de bases

DNA – Ácido desoxirribonucléico

rRNA - Sequencia de dna pertencente aos ribossomos

tRNA – Sequências de dna co funções de transporte de aminoácidos

NCBI – National Center for Biotechnology Information

FBN – Fixação Biológica de Nitrogênio

MoFe – Proteína Molibdênio-Ferro

GenBank – Banco de dados público do National Center for Biological Information

KASS – KEEG Automatic Annotation Server

KEGG – Kyoto Encyclopedia of Genes and Genomes

RAST – Rapid Annotation Programming Genefinding Algorithm

16S – rRNA 16S ribosomal Ribonucleic Acid (ácido ribonucleico ribosomal)

16S-23S-5S – Ribosomal operon (operon ribosomal)

FASTA – Formato utilizado para armazenar sequências de bases e de aminoácidos em arquivo texto

N2 – Nitrogênio

OLC - Overlap-layout-consensus

NGS – Sequênciamento de nova geração

ANI – Identidade Média Nucleotídea

PCR – Polymerase Chain Reaction

Primer – Oligonucleotídeo iniciador do processo de PCR

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Algumas bactérias possuem a capacidade de realizar associações com plantas e promover o crescimento vegetal através da fixação biológica de nitrogênio. A bactéria *Azoarcus olearius* DQS4 é uma das bactérias que possuem estas habilidades. A partir da publicação da descoberta desta bactéria, surgiu o interesse na biologia e em uma potencial aplicação econômica deste organismo. *A. olearius* DQS4 também possui enzimas capazes de degradar óleos e podem ter aplicações biotecnológicas tais como descontaminação de áreas biodegradadas. Um estudo mais aprofundado desta bactéria exige o sequenciamento de seu genoma assim como a montagem e anotação dos genes.

A realização da montagem e anotação do genoma da bactéria permitiu a identificação dos genes comuns a organismos filogeneticamente relacionados e os genes responsáveis pela fixação biológica de nitrogênio, assim como alguns genes de degradação de compostos aromáticos. A comparação do genoma de *A. olearius* DQS4 com o genoma de referência de *Azoarcus* BH72, assim como os demais genomas de bactérias do gênero *Azoarcus* disponibilizados no GenBank (*Azoarcus* KH32C, *Azoarcus* EBN1 e *Azoarcus toluclasticus*) é importante para definir as relações fisiológicas e filogenéticas entre esses organismos.

2. REVISÃO DE LITERATURA

2.1 FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

Existe um grande número de plantações de leguminosas, gramíneas e cereais no mundo com grande importância na produção de alimentos para a humanidade. Conforme DOBEREINER (1997), o elemento mais importante para elevadas produções na agricultura tropical é o nitrogênio. Ele esta presente em 80% da atmosfera na forma gasosa, mas as plantas não conseguem utilizá-lo. Existem bactérias que conseguem realizar a fixação do nitrogênio (N_2) através da sua conversão em amônia (NH_3), uma forma na qual as plantas conseguem captar. Esse processo é denominado fixação biológica de nitrogênio e é catalizado pelo complexo enzimático da nitrogenase. Esta associação entre plantas e bactérias é conhecida desde o início do século como simbiose das leguminosas, no qual as plantas são infectadas por bactérias capazes de realizar a fixação biológica de nitrogênio. (DOBEREINER *et al*, 1997). O uso desta técnica tem contribuído fortemente para colocar o Brasil em uma posição muito favorável na produção agrícola quando comparado aos demais países que fazem maior uso de fertilizantes no solo.

Segundo DOBEREINER *et al* (1997), um dos maiores desafios nesse campo seria encontrar bactérias que realizassem a fixação biológica de nitrogênio com plantas não leguminosas, principalmente gramíneas e cereais. Isto porque as mais importantes culturas do mundo são trigo, arroz e milho, que pertencem à família das *Gramineae* (HUREK e REINHOLD-HUREK, 2003). Uma das bactérias inicialmente encontradas com estas características foi do gênero *Azospirillum*, isoladas de cana-de-açúcar e cereais como milho, arroz e sorgo (BALDANI e DOBEREINER 1979). Com o passar dos anos foram sendo descobertas novas espécies de bactérias diazotróficas endofíticas, capazes de fixar nitrogênio e se multiplicar dentro de plantas sem lhes trazer malefícios, como duas espécies do gênero *Herbaspirillum* e uma de *Gluconacetobacter*, denominada *Gluconacetobacter diazotrophicus* (MUTHUKUMARASAMY, R. *et al* 2002). Bactérias isoladas do gênero *Azoarcus*, entre essas o *Azoarcus* sp. BH72, também são capazes de se associar endofiticamente com gramínes e fixar nitrogênio atmosférico (REINHOLD-HUREK *et*

al, 1993). Em 2013 CHEN e colaboradores publicaram a descoberta de uma nova bactéria do gênero *Azoarcus*, denominada de *Azoarcus olearius* DQS4. Segundo os autores, a bactéria apresenta grande semelhança com a estirpe BH72, justificando sua utilização como biofertilizante.

A partir desta descoberta surgiu o interesse na biologia da bactéria, conhecer o quanto ela é semelhante com as demais e verificar uma potencial aplicação econômica deste organismo. Segundo experimentos realizados pelo atutor, *A. olearius* é uma bactéria fixadora de nitrogênio capaz de formar associação endofítica com gramíneas de importância econômica e promover o crescimento vegetal (CHEN et al, 2013). Além disto, possui enzimas capazes de degradar óleos que podem ter aplicações biotecnológicas tais como descontaminação de áreas biodegradadas.

2.2 COMPLEXO ENZIMÁTICO DA NITROGENASE

Uma bactéria diazotrófica, fixadora de nitrogênio, deve possuir alguns genes necessários para realizar uma associação endofítica diazotrófica com plantas, e desta associação, trazer benefícios para a planta. Diferentes estudos ao longo dos anos estabeleceram que os genes envolvidos com o processo de fixação de nitrogênio através da nitrogenase *Nif* (ferro-molibdênio) seriam os genes *nifH*, *nifD*, *nifK*, *nifY*, *nifB*, *nifQ*, *nifE*, *nifN*, *nifX*, *nifU*, *nifS*, *nifV*, *nifW* e *nifZ* (ZHENG *et al*, 1998; METTICK e EDWARDS, 1995; DIXON e KAHN, 2004; HU *et al*, 2007; RUBIO e LUDDEN, 2008) citado por M. Aalves e C. Gehlen, (2011). Entretanto, DOS SANTOS e colaboradores (2012), baseados em uma comparação computacional entre várias espécies de bactérias diazotróficas, propôs um critério de avalização para indicar se uma dada bactéria é fixadora de nitrogênio. Segundo os autores, a bactéria deve possuir em seu genoma no mínimo os genes *nifH*, *nifD*, *nifK*, *nifE*, *nifN* e *nifB*. Os genes *nifHDK* codificam o complexo da nitrogenase dependente de molibdênio, onde o *NifH* codifica as subunidades da proteina *Fe* e o *NifD* e *NifK* codificam a subunidade alfa e beta da proteína *MoFe*. Os demais genes codificam o *FeMoco* e demais proteínas envolvidas na fixação biológica de nitrogênio (SEEFELDT *et al*, 2009) (FIGURA 1).

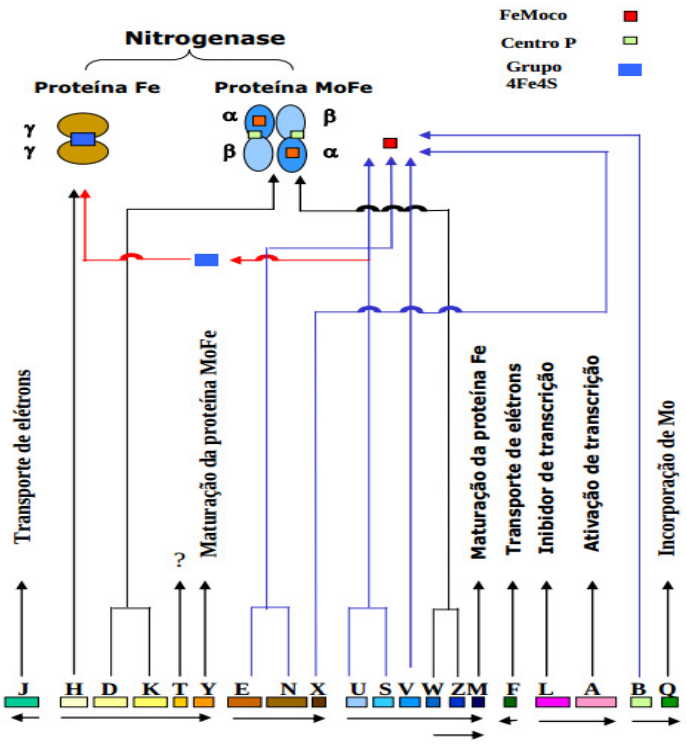

FIGURA 1: NITROGENASE NIF

FONTE: Gisele Klassen.

2.3 GÊNERO AZOARCUS

O gênero *Azoarcus* foi primeiramente proposto por REINHOLD-HUREK e colaboradores (1993), através da publicação da descoberta da espécie *Azoarcus indigens*. Este gênero pertence à família *Rhodocyclaceae*, ordem *Rhodocyclales*, classe *Betaproteobacteria* (REINHOLD-HUREK E HUREK, 2006). São 8 as espécies conhecidas do gênero *Azoarcus*: *Azoarcus indigens* (REINHOLD-HUREK *et al*., 1993), *Azoarcus communis* (REINHOLD-HUREK *et al.,* 1993), *Azoarcus tolulyticus* (ZHOU *et al.,* 1995), *Azoarcus evansii* (ANDERS *et al.,* 1995), *Azoarcus anaerobius* (SPRINGER *et al.,* 1998), *Azoarcus toluclasticus* (SONG *et al.,* 1999), *Azoarcus toluvorans* (SONG *et al.,* 1999) e *Azoarcus buckelii* (MECHICHI *et al.,* 2002). As bactérias do gênero *Azoarcus* foram encontradas em caules, raízes, sedimentos aquíferos, solos contaminados com petróleo e lamas de esgoto. Membros deste gênero possuem um metabolismo estritamente aeróbio, com exceção do *Azoarcus anaerobius*. Algumas espécies fixam nitrogênio, mas exigem condições

microaeróbicas. As espécies do gênero já estudadas possuem conteúdo G+C em seu genoma, entre 62 a 68 % (REINHOLD-HUREK e HUREK, 2006).

O *Azoarcus* sp. BH72 é uma bactéria diazotrófica endofítica fixadora de nitrogênio, capaz de colonizar o interior de raízes de plantas como o arroz e outras gramíneas. Esta estirpe foi isolada a partir da superfície esterilizada de raízes de uma gramínea *Kallar* de um solo de baixa fertilidade em Punjab no Paquistão (HUREK e REINHOLD-HUREK, 2003). A mesma pertence ao *Filo Proteobacteria* e foi descrito através de duas espécies válidas, *Azoarcus communis* e *Azoarcus indigens*, mas seus estudos de hibridização DNA-DNA mostraram que ela é diferente destas espécies. Devido à falta de mais estirpes deste gênero para testes, ela permaneceu sem definição de espécie (REINHOLD-HUREK *et al*., 1993b).

O genoma de referência de montagem do Azoarcus sp. BH72 foi o genoma da bacteria *Azoarcus Ebn1*. A comparação genômica entre estas bactérias possui baixo grau de sintenia. O genoma da bactéria BH72 é formado por um chromossomo circular com conteúdo GC de 67,92%, possuindo 4,376,040 pares de base e codificando 3,992 proteínas. Possui sequências codificas envolvendo síntese de componentes de superfície potencialmente importantes para a interação micróbrio-planta. Não é uma bactéria patógena, apenas possui algumas enzimas de degradação de parede celular. A estirpe analisada encontra-se adaptada e relativamente estável em relação ao ambiente que foi encontrada apresentando baixo nível de expressão gênica(Krause, A *et al., 2006*).O *Azoarcus sp*. KH32C é uma bactéria desnitrificadora (NISHIZAWA *et al*., 2012). Foi isolada de uma amostra de solo de um campo agrícola onde arrozais e lavouras de soja foram cultivadas em rotação de cultura a cada 2 anos (TAGO *et al*., 2011). Esta bactéria também possui grande habilidade em reduzir N_2O para amônia, indicando um potencial uso desta bactéria para redução da emissão de óxido nitroso dos campos agrícolas (NISHIZAWA *el al.,* 2012). O genoma da estirpe KH32C é composta de um cromossomo circular de 5 Mb e um plasmídeo circular de 737.589 bp com uma média de 65,1% e 64,5% de conteúdo G+C respectivamente. Foi identificado tanto em seu cromossomo quanto em seu plasmídeo o conjunto de genes desnitrificadores. Além disto, em seu cromossomo, foram encontrados também genes relacionados com a fixação de nitrogênio, assim como genes para biossíntese de polissacarídeos (NISHIZAWA *el al*., 2012), que podem estar envolvidos na

comunicação planta-microorganismo e simbiose (PARADA *et al*., 2006). Também foram identificados genes relacionados à fixação de carbono em seu plasmídeo. Os resultados sugerem que *A.* KH32C pode ter a capacidade de crescer de forma autotrófica, para estabelecer associação simbiótica com as plantas e degradar compostos aromáticos (NISHIZAWA *et al*., 2012)

O *Azoarcus sp.* estirpe EbN1 é uma bactéria desnitrificadora anaeróbica, capaz de metabolizar vários componentes aromáticos incluindo hidrocarbonos. Esta estirpe é relativamente próxima a muitas espécies do gênero *Thauera*, ao contrário de espécies de *Azoarcus*, que são encontrados em associação simbiótica com plantas. Esta estipre foi isolada de solo e de água doce. O genoma da estirpe EbN1 é composto de um cromossomo circular de 4.3 Mb e dois plasmídeos de 0.21 e 0.22 Mb codificando 4.603 proteínas. A ausência de genes necessários para a fixação de nitrogênio e interação com plantas diferencia esta estirpe ecofisiologicamente do grupo dos *Azoarcus* simbiontes conhecidos (RABUS *et al*., 2005).

O *Azoarcus toluclasticus* estirpe ATCC 700605 é uma bactéria desnitrificadora de nitrogênio e degradadora de componentes aromáticos (SONG *et al*., 1999). Foi isolada de um sedimento aquífero de pouca profundidade em *Moffett Field* no estado da Califórnia, EUA. A análise de seu gene 16S rRNA possui similariedade superior a 99% quando comparado com *Azoarcus tolulyticus.* Segundo dados do autor, em uma análise filogenética entre bactérias degradadoras de componentes aromáticos e desnitrificadoras, esta estirpe fica localizada entre os grupos de *Azoarcus tolulyticus* e *Azoarcus evansii*. Também foi identificado que possui a forma de bastonete com flagelos de motilidade curtos e podem crescer em acetato, benzoato, Pir-uvate, succinato, D-xilose, L-arabinose, D-ribose, D- galactose, sacarose, lactose, maltose, adipato, lactato, manitol, aspartato, prolina ou arginina, sob condições aeróbicas e desnitrificantes (SONG e HÄGGBLOM, 1999).

2.4 *AZOARCUS OLEARIUS* DQS4

Uma nova espécie do gênero *Azoarcus* foi descrita recentemente e denominada *Azoarcus olearius* DQS4. Ela foi encontrada e isolada de uma amostra de solo contaminado com óleo em uma refinaria de petróleo na cidade de Kaoshiung em Taiwan. Foi constatado por CHEN e colaboradores (2013), que ela possui

metabolismo estritamente aeróbico e é uma bactéria fixadora de nitrogênio. Além da capacidade de fixar nitrogênio, esta bactéria é capaz de degradar óleos e de colonizar plantas de arroz endofiticamente (JAMES E., comunicação pessoal). Devido a estas características, esta bactéria é potencialmente importante para aplicações biotecnológicas.

Análises fenotípicas e genotípicas mostraram que *Azoarcus indigens* LMG 9092T e *Azoarcus communis* DSM 12120T são filogeneticamente relacionados *a A. olearius* DQS4 (CHEN et al, 2013). A comparação de sequências baseada no gene 16S rRNA mostrou uma identidade de 99,9% com o gene 16S rRNA do *Azoarcus* sp. BH72, 97,4% do *Azoarcus indigens* LMG 9092T e 96,4% do *Azoarcus communis* DSM 12120T (FIGURA 2). Também foi determinada a similaridade de parte da sequência do gene *nifH* (FIGURA 3), que mostrou que a estirpe DQS4 está intimamente relacionada com o *Azoarcus* sp. BH72 (99,4 %) e menos relacionada com o *Azoarcus indigens* VB32T (95,5 %), *Azoarcus communis* SWub3T (87,1 %), *Azoarcus communis* S2 (85,8 %) e *Azoarcus tolulyticus* Td-1 (79,5 %) (CHEN *et al*., 2013).

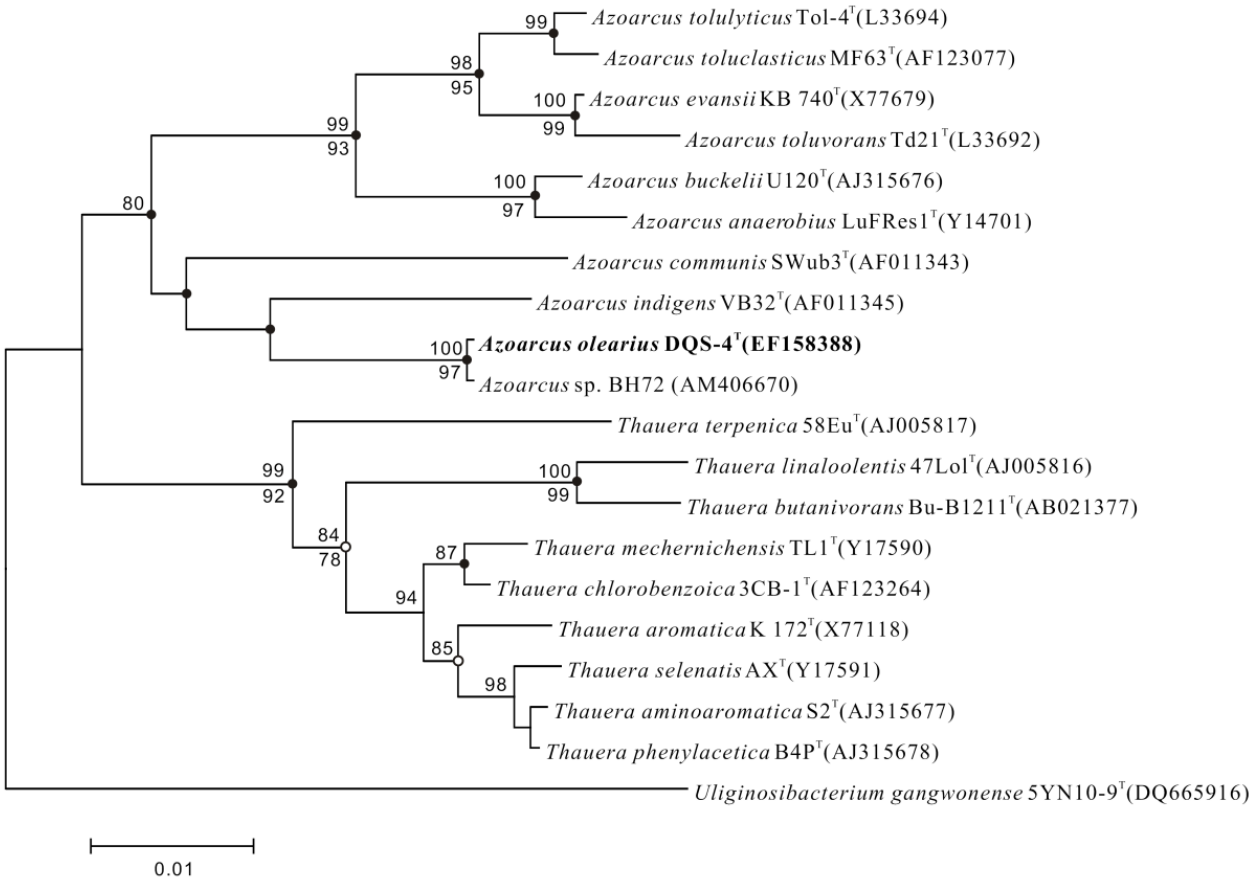

FIGURA 2 ÁRVORE FILOGENÉTICA BASEADA NO GENE 16S PARA O GÊNERO *Azoarcus*
Árvore filogenética baseada em seqüências do gene 16S rRNA mostrando a posição do *Azoarcus olearius* DQS4 relacionados com as demais espécies da classe B-*Proteobacteria*.

FONTE: Chen et al, (2013).

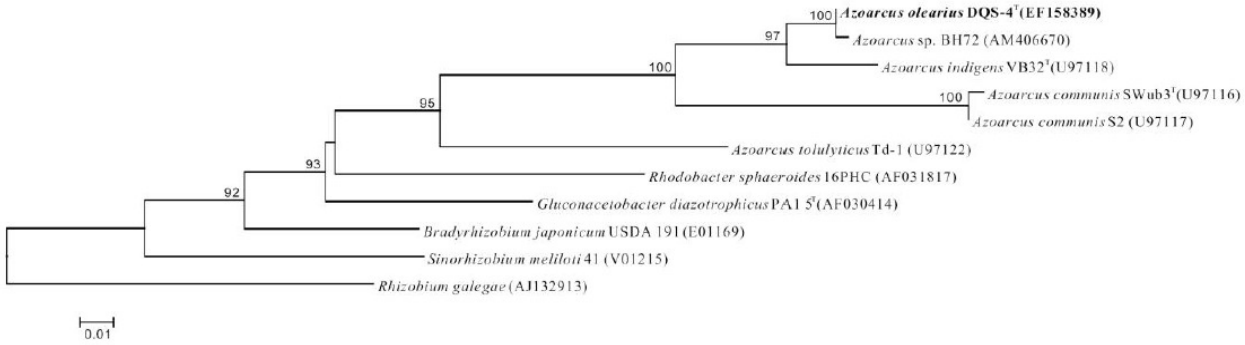

FIGURA 3 ÁRVORE FILOGENETICA BASEADA NO GENE *nifH*
Árvore filogenética baseada em sequências do gene *nifH* mostrando as relações entre o *Azoarcus olearius* DQS4 e outras bactérias fixadoras de nitrogênio.

FONTE: Chen et al, (2013).

Com base em dados genotípicos e fenotípicos e nas comparações das sequências dos genes 16S rRNA e *nif*H se conclui que DQS4 ocupa uma posição distinta dentro do gênero *Azoarcus* e, portanto, foi sugerido que representa uma nova espécie. O nome proposto por CHEN e colaboradores (2013), foi *Azoarcus olearius*. O nome *olearius* (de pertencer ao óleo) descreve o ambiente de onde a estirpe foi isolada.

2.5 MONTAGEM DE GENOMAS

O sequenciamento de um genoma é a primeira etapa do processo e existem diferentes tipos de sequenciadores que utilizam diferentes tipos de tecnologias de sequenciamento. O objetivo é assegurar que nenhuma região do genoma fique sem representação nos fragmentos, evitando a ausência de sobreposições de alguma parte do genoma. O sequenciamento gera leituras (*reads*) a partir de bibliotecas genômicas. Os montadores interpretam as leituras e realizam o agrupamento das mesmas em *contigs*, através das sobreposições das mesmas. Em uma nova etapa, é realizado o agrupamento e extensão dos *contigs* gerando *scaffolds*. Posteriormente, após o fechamento de regiões não contíguas (*gaps*), de baixa ou sem cobertura, é obtido o genoma final do organismo. (FIGURA 4)

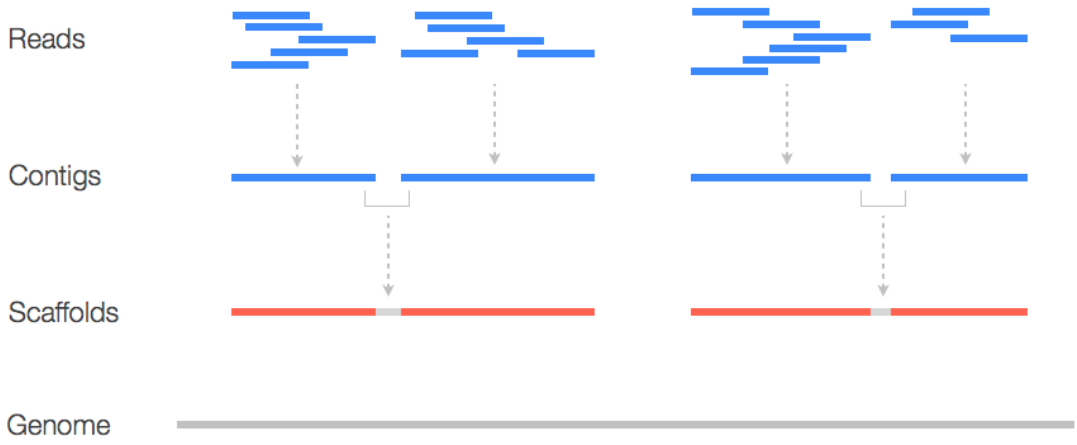

FIGURA 4 PROCESSO DE MONTAGEM DE UM GENOMA
Fonte: Dísponível em <http://ecoevo.unit.oist.jp/lab/?page_id=141>

Existem duas estratégias de montagem, uma em que as leituras podem ser ordenadas com base em um genoma de referência e a estratégia de montagem *De Novo.* Esta ultima estratégia é a mais utilizada, pois monta o genoma sem as informações do genoma de referência. Atualmente as duas estratégias são utilizadas em conjunto, após a montagem *De Novo,* utiliza-se as leituras novamente para realizar a ordenação dos *contigs* e *scaffolds*. Assim não há influência nas características do genoma. Dentre os diferentes montadores utilizados nas montagens, existem duas abordagens mais conhecidas, o algoritmo de montagem OLC e a utilização de algoritmos que fazem o uso de grafos de *De Brujin*. A abordagem OLC, realiza o alinhamento de sobreposições entre todos os *reads* gerando uma sequência consenso, possui uma flexibilidade com o tamanho dos reads e uma redução na taxa de erros. Alguns exemplos de montadores que utilizam esta abordagem são o *Celera Assembler* (MILLER et al, 2008) e o *NewBler Assembler* (Roche).

As abordagems com algoritmos de grafos de *De Brujin* trabalham com valores de *k-mer* e são utilizados pelos montadores de segunda geração que produzem leituras curtas (25 a 100 pb). Um dos objetivos da utilização desta abordagem é de reduzir o custo de processamento computacional, tornando a montagem mais rápida e eficiente, onde um grande número de sobreposições não precisa ser processado. Um *k-mer* é o valor mínimo de sobreposição utilizado para gerar os grafos de *De Brujin*, e ele é fixado em um valor de tamanho *k* para todas as sub-leituras geradas a partir das leituras originais. Cada sub-leitura possui um número *k* de bases onde são geradas sobreposições entre as mesmas a fim de se produzir um *draft* (rascunho) do genoma a ser montado. Na grande maioria das montagens utilizando grafos *De brujin*, é recomendada a utilização de bibliotecas de leituras pareadas. Esta informação torna possível a ligação dos *contigs* e *scaffolds* que estão separados por lacunas que não obtiveram sobreposições, estas chamadas de *gaps*. Alguns exemplos de montadores que utilizam este algoritmo são o EULER-SR (CHAISSON e PEVZNER, 2008), SOAPdenovo2 (LUO *et al*, 2012) e o Velvet (ZERBINO e BIRNEY, 2008).

As etapas finais de uma montagem são a ordenação do genoma e o fechamento dos *gaps*. Atualmente existem alguns programas que realizam esta tarefa como o FGAP (PIRO *et al*, 2014).

Após a montagem do genoma é realizado a anotação, processo que tem por objetivo identificar os genes e, quando possível, atribuir funções às proteínas codificadas no genoma assim como a identificação dos rRNAs e tRNAs. A identificação dos genes normalmente é realizada por algoritmos que utilizam o programa BLAST (ALTSCHUL *et al*., 1997) buscando por similariedade dos genes desejados com os banco de dados de genomas disponíveis. Algumas plataformas que realizam anotação são a plataforma RAST (AZIZ *et al*., 2008) e o programa SILA (VIALLE *et al*, 2013).

3. OBJETIVOS

3.1 OBJETIVO GERAL

• Montagem e anotação do genoma parcial da bactéria *Azoarcus olearius* DQS4.

3.2 OBJETIVOS ESPECÍFICOS

• Realizar a montagem do genoma parcial da bactéria *Azoarcus olearius* DQS4.

• Obter um mapa físico dos *contig*s genômicos de *Azoarcus olearius* DQS4 utilizando um genoma de referência.

• Realizar a anotação parcial do genoma e procurar pela existência de genes de interesse.

• Comparar o genoma do *Azoarcus* DQS4 com outros genomas de bactérias fixadoras de N2 que fazem associação com plantas endofíticamente.

• Comprovar se o *Azoarcus* DQS4 realmente é uma bactéria endófita.

4. MATERIAIS E MÉTODOS

4.1 FERRAMENTAS UTILIZADAS

4.1.1 CLC *Genomic Workbench*

O *CLC Genomic Workbench* (KNUDSEN e FLENSBORG, 2008) é um programa utilizado para análises de sequências de DNA, RNA e proteínas e montagens de pequenos genomas. Também é utilizado para análise de expressão gênica, modelagem de primers, clonagem, análises filogenéticas e gerenciamento de sequências genéticas entre várias outras características. Difere-se dos demais programas pela sua intuitiva e amigável interface gráfica com o usuário. Realiza análise de qualidade nos dados brutos utilizando os valores de PHRED (EWING et al, 1998). O programa foi utilizado na realização das montagens iniciais, comparações de sequências e análises de qualidade das leituras assim como das montagens finais.

4.1.2 PHRED

Os valores de PHRED (EWING et al, 1998), são utilizados para estimar a probabilidade de erro de cada base em uma sequência. Valores acima de 20 são considerados bons resultados, onde o valor “10”, por exemplo, expressa a probabilidade de erro de 1 em 10. O valor de “20”, expressa a probabilidade de 1 em 100 e assim sucessivamente.

4.1.3 SOAPdenovo2

O SOAPdenovo2 (LUO et al, 2012), é um montador que realiza montagens de estratégia de novo utilizando sequenciamentos de nova geração, NGS, de leituras curtas. Possui um bom desempenho com regiões de repetições de montagem,

proporcionando melhor cobertura e gerando scaffolds maiores. Possui um melhoramento na função de fechamentos de gaps e montagens de genomas grandes. Possui seis módulos de correção de erros de leituras e faz o uso de algoritmos de montagem baseados em grafos de *De Brujin*. O programa foi utilizado na realização de algumas montagens que foram utilizadas posteriormente também como conjunto de dados para o fechamento de gaps.

4.1.4 Velvet

O Velvet (ZERBINO, 2008), é um montador baseado em grafos de *De Brujin*, ele é um software de código aberto e pode ser utilizado gratuitamente. É dividído em dois módulos de execução, Velveth e Velvetg. Permite realizar testes de montagens com vários números de k-mers a fim de se verificar qual dos valores de k-mers é o mais adequado para determinado dados de origem. O velvet foi um dos primeiros montadors a trabalhar com leituras curtas. É um excelente montador para se realizar montagens *De Novo*. O programa foi um dos primeiros a ser utilizando, além de vários testes de montagens, o mesmo foi utilizado na realização de algumas montagens que foram utilizadas posteriormente também como conjunto de dados para o fechamento de gaps.

4.1.5 Newbler Assembler

NewBler Assembler (Roche), é um software de montagem distribuído juntamente com o sequenciador 454 da Roche. É considerado um excelente montador de pirossequenciamento devido a prever os erros deste tipo de sequenciamento gerando contigs mais precisos. Conhecido como GS *De Novo* Assembler, identifica sobreposições entre os reads construindo alinhamentos múltiplos de sobreposição. É um montador que gera montagens muito boas para genomas de bactérias e para leituras curtas, de no máximo 500 pb. Após melhorias, passou a trabalhar com montagens híbridas onde podem ser combinadas leituras curtas, provenientes do sequenciador 454 e leituras longas, normalmente de

sequenciamentos antigos do tipo Sanger em uma mesma montagem. (CHAISSON e PEVZNER, 2008). O programa foi utilizado em algumas montagens e na montagem final onde foram reunidos vários tipos de montagens e leituras de tamanhos diferentes para a conclusão do genoma consenso.

4.1.6 MUMmer / Prommer

O MUMmer (KURTZ et al., 2004) é um programa que realiza o alinhamento e comparação entre genomas. O pacote utiliza o gnuplot para a geração dos dotplots gráficos para a visualização das comparações. Esta disponível em <http://mummer.sourceforge.net/>. O programa foi utilizado para realizar o alinhamento e comparação entre o *draft* genomico do A. DQS4t e o genoma de referência, assim como para os demais genomas utilizados nas comparações. Todos os dotplots gráficos de comparações foram gerados com este programa.

4.1.7 FGAP

O programa FGAP (PIRO et al, 2014) é um programa especializado na finalização de genomas. Utiliza um conjunto de dados de várias versões de montagens que não foram bem sucedidas no processo de montagem do genoma. Com este conjuto de dados, realiza a análise de similariedade destas regiões assim como as extremidades dos contigs a fim de encontrar a melhor sequência que se encaixa no gap alvo. O programa FGAP faz uso do programa BLAST para realizar a busca por similaridade. O programa foi utilizado para realizar o fechamento automático dos gaps remanescentes do genoma, o qual gerou a versão final do genoma.

4.1.8 RNAmmer

O programa RNAmmer (LAGESEN et al, 2007) faz a busca pela sequência dos operons ribossomais em genomas, trazendo como resultado a posição dos mesmos na sequência submetida. Está disponível no endereço eletrônico <http://www.cbs.dtu.dk/services/RNAmmer/>. O programa foi utilizado na localização das sequências dos operons ribossomais.

4.1.9 Artemis

O programa Artemis (CARVER et al, 2008), é um programa desenvolvido em linguagem Java, capaz de ser executado em qualquer sistema operacional com suporte a esta linguagem. Tem por objetivo permitir a visualizaçõa e anotação de genomas além de permitir ao usuário inserir anotações manualmente. É utilizado para realizar revisões manuais em anotações automáticas e facilitar o estudo detalhado dos genes anotados. O programa foi utilizado para revisar as anotações automáticas.

4.1.10 ACT

O programa ACT (CARVER et al, 2008), é um programa semelhante ao Artemis, mas permite a visualização de dois genomas após comparação via BLASTn. É possível realizar anotações assim como no programa Artemis. O programa foi utilizado na comparação e análise entre o genoma final e o genoma de referência assim como na geração das figuras comparativas de regiões de inserção entre os dois genomas.

4.1.11 BLAST

O programa BLAST (ALTSCHUL et al., 1997), faz comparações de similariedade entre sequências de nucleotídeos e aminoácidos. Tem como resultados das buscas os valores de identidade e similariedade entre as sequências

analisadas. O programa foi utilizado nas comparações de similariedade de sequências de nucleotídeos ou de proteínas entre os genomas do gênero.

4.1.12 SILA

O programa SILA (VIALLE et al. 2013) é um programa que realiza a anotação de genomas e esta disponível no endereço eletrônico <http://200.236.3.34/SILA/login.jsp>. O programa foi um dos utilizados no processo de anotação automática do genoma.

4.1.13 RAST

A plataforma RAST (AZIZ et al., 2008), do inglês *Rapid Annotation using Subsystem Technology* oferece o serviço de anotação automática de genomas de bactérias e archeas realizando anotações de alta confiabilidade. O programa foi um dos utilizados no processo de anotação automática do genoma.

4.1.14 Identidade Nucleotídica Média (ANI)

O programa ANI *calculator* calcula a estimativa da identidade média nucleotídea através da comparação entre dois genomas. Tipicamente, os valores entre genomas da mesma espécie possuem mais de 95% de similaridade (GORIS et al, 2007). Valores abaixo de 75% não são confiáveis podendo ser analisados pelo programa AAI, que realiza a comparação de aminoácidos. Estas ferramentas suportam tanto genomas completos como rescunhos em fasta ou multi-fasta. Esta disponível na internet, através do endereço eletrônico <http://enve-omics.ce.gatech.edu/ani/>. O programa foi utilizado nas comparações de estimativa de identidade média nucleotídea entre os genomas analisados neste trabalho.

4.1.15 KEGG / KASS

O KEEG (*Kyoto Encyclopedia of Genes and Genomes*) é um recurso de banco de dados para compreensão de alto nível dos sistemas biológicos como celular, de organismos e ecossistemas a partir de informações de nível molecular, especialmente conjuntos de dados moleculares em larga escala gerados pelo sequenciamento de genomas e outras tecnologias experimentias de alto rendimento. O KASS (KEEG *Automatic Annotation Server*) é um sistema para a anotação funcional de genes. Faz uso da ferramenta BLAST ou GHOST para comparar um gene contra um conjunto de sequências referenciais no banco de dados do KEGG GENES. O resultado da comparação monta um (KEGG *Orthology*) gerado automaticamente pelo KEEG, que é um mapa de uma via metabólica. (MORIYA, Y et al., 2007). Este programa foi utilizado para a geração dos mapas das vias metabólicas de Nitrogenio e Enxofre na comparação entre o A. DQS4 e A. BH72.

4.1.16 VIM

O VIM é um programa de edição de texto mantido como *software* livre que acompanha a maioria das distribuições Linux, atualmente também está disponível para outros sistemas operacionais. Além de possuir várias funções úteis como buscas e substituições, também permite o uso de expressões regulares para especificar regiões onde vários comandos devem ser executados.

O programa VIM foi utilizado largamente como editor de texto em ambiente Unix e teve como tarefa mais importante a formatação dos cabeçalhos das bibliotecas originais de sequenciamento, que permitiram com que o programa de montagem Newbler identificasse os reads pareados. Foram utilizados comandos que realizaram a remoção dos espaçamentos no nome dos cabeçalhos dos arquivos, substituindo-os por underlines. Os comandos encontram-se no apêndice deste trabalho juntamente com os scripts utilizados.

4.1.17 Scripts desenvolvidos e utilizados

Foram utilizados três tipos de scripts para resolver situações específicas encontradas no decorrer da montagem do genoma do *Azoarcus olearius* DQS4.

O “Script de quebra em 500 pb”, foi desenvolvido pelo doutorando do programa de ciências biológicas da UFPR, Rodrigo Cardoso (não publicado). Como o programa de montagem Newbler trabalha com leituras curtas, de no máximo 500 pb, foi necessário o uso deste script para realizar a fragmentação de contigs longos. Este script fragmenta os contigs em tamanhos de 500 pb, mantendo uma sobreposição de 50 bases em cada leitura para permitir a remontagem dos contigs originais. O script foi desenvolvido na linguagem de programação Phyton e esta disponível no apêndice deste trabalho.

O “Script pré-Newbler” é um script disponível na internet em <https://contig.wordpress.com/>, deve ser rodado nos arquivos de leituras pareadas, os quais o montador Newbler não consegue identificar os pares. O Newbler é um montador que trabalha com reads provenientes do sequenciador 454 da Roche e não trabalha com leituras pareadas, assim, através do uso deste sript, o montador passa a identificar a existência dos pares permitindo a utilização de leituras do sequenciador Illumina® nas montagens. Este script vem a substituir o uso dos comandos no editor de texto VIM, mencionado anteriormente, sendo que ambos realizam a mesma tarefa. O conhecimento da existência deste script ocorreu na metade do processo onde em parte dos dados foi utilizado o programa VIM e o restante este script. O script foi elaborado na linguagem AWK, usada para deixar os scripts de *shell* com mais recursos. Para utiliza-lo, deve-se substiruir em “arquivo_de_reads.fasta” pelo nome do arquivo de entrada e renomear o nome do arquivo de saída caso seja necessário.

O “Script gera DATASET de retirada de N”, desenvolvido pelo próprio autor em linguagem Matlab, da *The mathworks*, foi utilizado para remover as bases ambíguas, bases “N”, do conjunto de dados submetido ao FGAP (PIRO *et al*, 2014). O script foi desenvolvido devido às restrições do programa FGAP de não aceitar montagens com bases ambíguas como dado de entrada para o fechamento dos gaps. O script faz primeiramente o agrupamento das bases “N” em uma única base “N” e

posteriormente ela é removida repartindo o contig em dois novos contigs e assim sucetivamente.

4.2 GENOMAS UTILIZADOS NA COMPARAÇÃO

Foram realizadas buscas nos bancos de dados de genomas públicos a fim de se encontrar todos os genomas sequenciados e montados existentes pertencentes ao gênero *Azoarcus*. Apenas os genomas apresentados na (TABELA 1) foram encontrados até a data das análises.

TABELA 1 GENOMAS DE REFERÊNCIA PARA VALIDAÇÕES DOS DADOS

	Estirpes						
Características	**BH72**	**KH32C**		**EBN1**			**ATCC700605**
Acesso Genbank	AM406670	AP012304		CR555306			ARJX01000000
	Chromosome	Chromosome	Plasmid	Chromosome	Plasmid 1	Plasmid 2	Chromosome
Tamanho (pb)	4,376,040	5,081,166	737,589	4,296,230	207,355	223,67	5,924,508
GC (%)	67.92	65.10	64.5	65.12	57.63	63.11	66.00
# ORFs	4,036	4,531	657	4,133	274	196	5,076
# rRNA operons	4	5	-	4	-	-	3
# tRNA	56	64	-	58	-	-	50

FONTE: National Center for Biotechnology Information, U.S. National Library of Medicine. 2013.

4.3 FLUXOGRAMA DE ATIVIDADES REALIZADAS

As FIGURAS 5 e 6 demonstram a sequência de atividades elaboradas no decorrer da análise, montagem e anotação do genoma do *Azoarcus olearius* DQS4. Uma descrição detalhada de todo o processo segue no tópico 5.3 deste trabalho. Os números em círculos no fluxograma representam atalhos, como se fossem túneis, onde uma seta para um dado número representa que o fluxo segue para a próxima aparição no decorrer do fluxograma, onde o fluxo continuará para uma próxima ação.

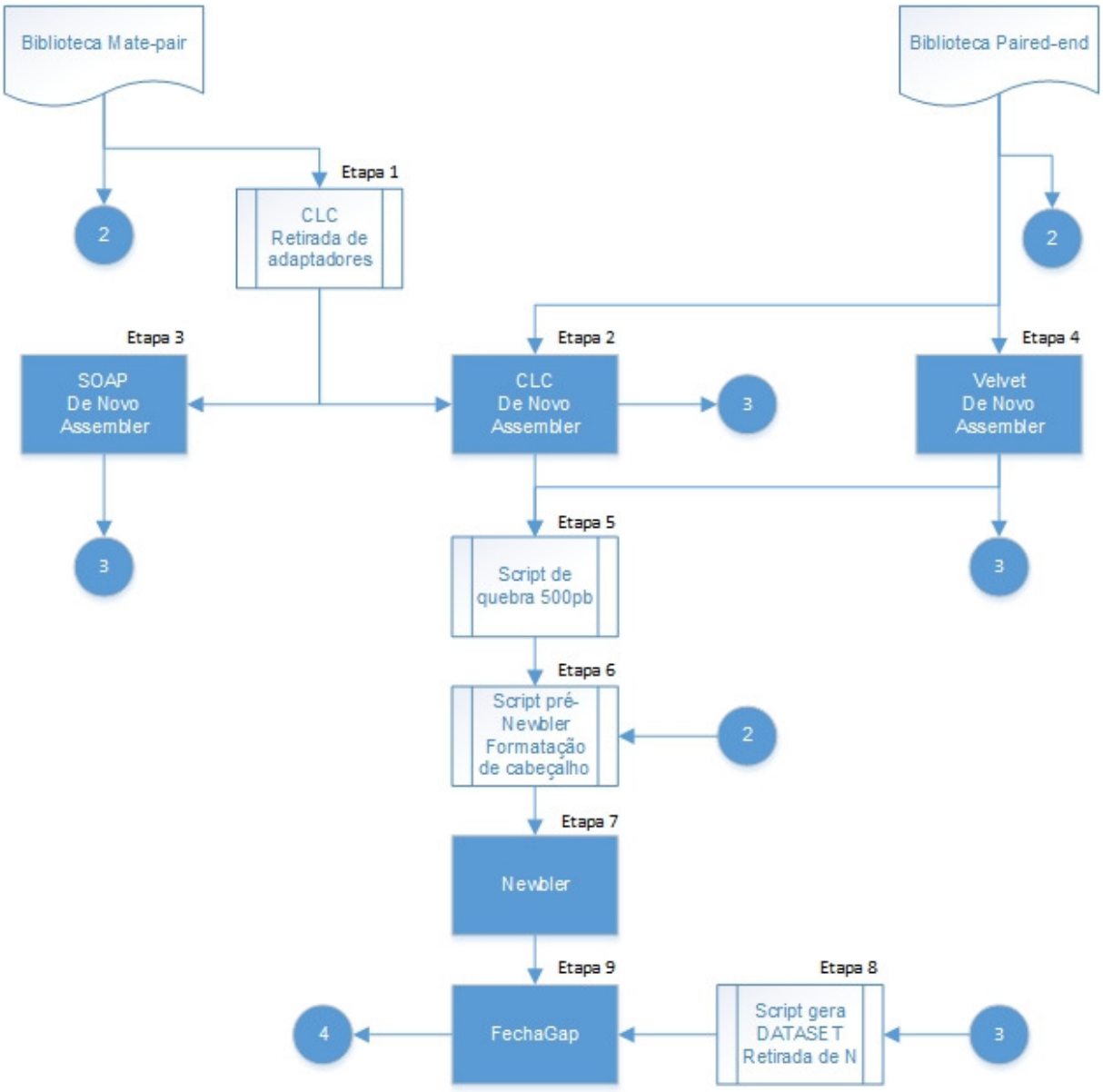

FIGURA 5 PIPELINE DE ATIVIDADES
FONTE: O próprio autor

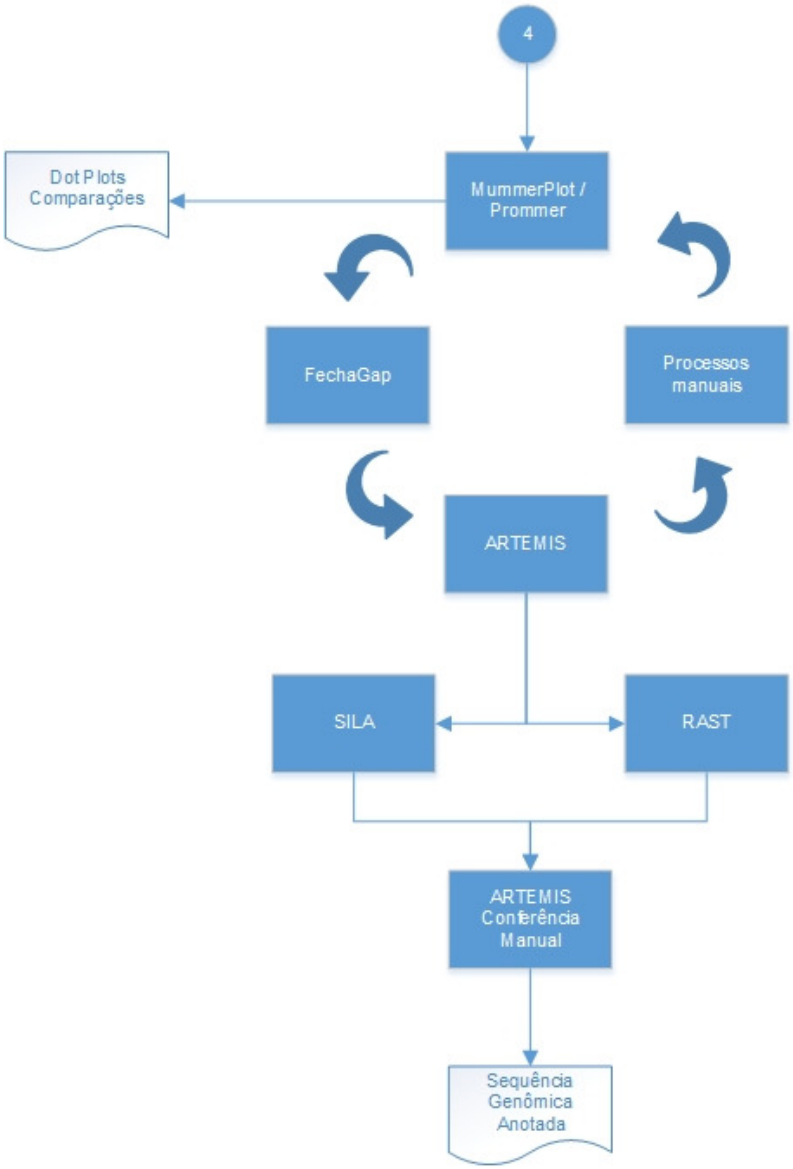

FIGURA 6 PIPELINE DE ATIVIDADES

FONTE: O próprio autor

5. RESULTADOS

5.1 ORIGEM DOS DADOS

A. olearius DQS4 foi cedido por Euan James, do *John Hutton Institute*, Reino Unido. O sequenciamento do genoma da bactéria *A. olearius* DQS-4 T foi realizado no Núcleo de Fixação de Nitrogênio, do Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná, utilizando o equipamento Illumina® MiSeq. Foram realizados dois sequenciamentos: um utilizando a estratégia de pareamentos longos (*mate-pair*) e outro utilizando pareamentos curtos (*paired-end*). Foram produzidas duas bibliotecas de leituras divididas em quatro arquivos (TABELA 2).

TABELA 2 ESTATÍSTICAS DE SEQUENCIAMENTO DO GENOMA DE A. OLEARIUS DQS4

Nome	Nº Reads	Tamanho Médio (pb)	Tamanho Máximo (pb)	Tamanho de fragmento (pb)
Azoarcus_matepair_R1.fastq	3.803.312	127,2	254	500 – 12500
Azoarcus_matepair_R2.fastq	3.803.907	132,1	254	
Azoarcus_pairend_R1.fastq	2.305.239	164,7	234	500 - 2500
Azoarcus_pairend_R2.fastq	2.305.239	166,1	234	

FONTE: O próprio autor.

A cobertura das leituras foi calculada em relação ao genoma de referência de *Azoarcus* BH72 e indica o número de vezes que o genoma está representado no sequenciamento. O valor de cobertura obtido na montagem com a biblioteca *Mate-Pair* foi calculado em 225,38 vezes e para a biblioteca *Paired-end* em 174,26 vezes.

5.2 ANÁLISE DOS DADOS BRUTOS

A análise dos dados brutos produzidos no sequenciamento foi realizada com auxilio do programa CLC *Genomic Workbench*. Nesta etapa foi possível verificar as características das leituras, em formato “fastq”, e suas respectivas qualidades. As

duas bibliotecas foram importadas no programa de maneira pareada, especificando-se a distância entre os pares. A grande maioria das sequências apresentou um tamanho de 250 pares de bases (FIGURA 7). O conteúdo GC% das leituras ficou com uma média próxima a 68% (FIGURA 8) e a qualidade média de valor Phred igual a 37 (FIGURA 9). O programa CLC possui uma ferramenta de *tuning* de qualidade, *Trim Sequences,* esta ferramenta tem por objetivo realizar cortes nas leituras, removendo tanto adaptadores de sequenciamento como sequências específicas, por exemplo, bases ambíguas, bases de baixa qualidade, uma quantidade desejada de bases nos finais das leituras assim como remove leituras muito curtas ou parte de leituras muito grandes com a finalidade de padronização. Este filtro foi aplicado para eliminar as bases de baixa qualidade nas leituras, mas não foi utilizado no pipeline de atividades. Foi verificada uma redução no tamanho das sequências após a utilização do filtro, sendo que os resultados das montagens pioravam com uso das sequências filtradas.

(A) Mate-pair (B) Paired-end

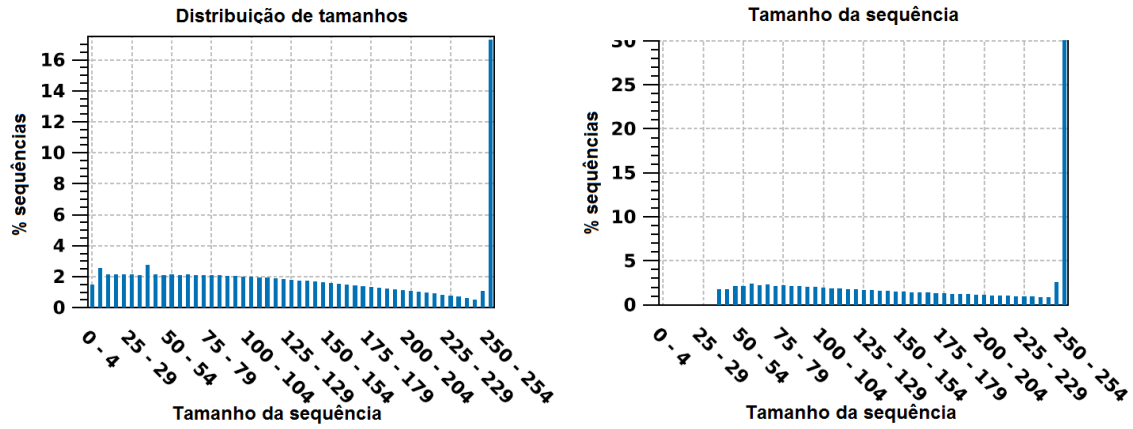

FIGURA 7 DISTRIBUIÇÃO DE TAMANHO DAS LEITURAS

O eixo X representa o tamanho das leituras em pares de base e o eixo Y é a porcentagem de leituras com o determinado tamanho.

FONTE: O próprio autor através do programa CLC Genomics WorkBench

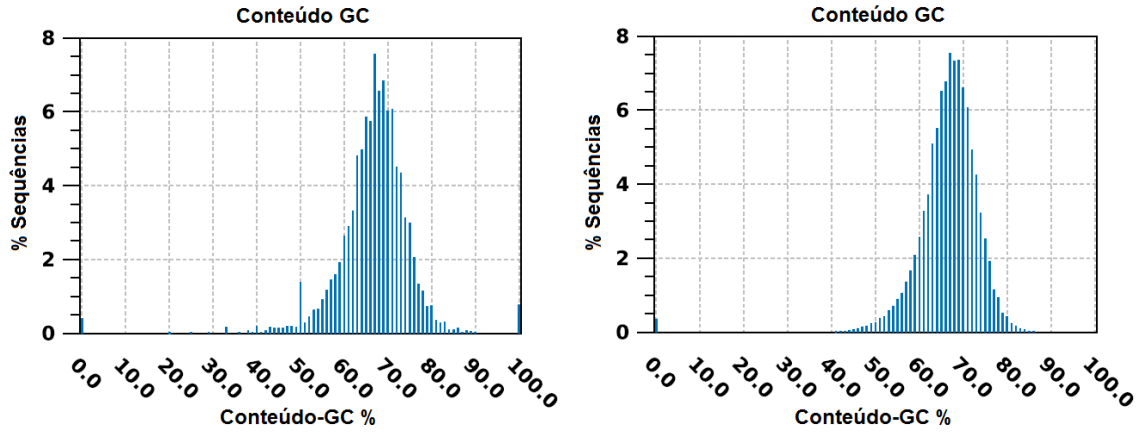

FIGURA 8 DISTRIBUIÇÃO DO CONTEÚDO GC DAS LEITURAS

Calculado com o numero de bases G e C comparado com todas as bases, incluindo bases ambíguas, caso existam. O eixo X representa o conteúdo GC relativo das leituras em porcentagem e o eixo Y é a porcentagem de leituras com o determinado valor.

FONTE: O próprio autor através do programa CLC Genomics WorkBench

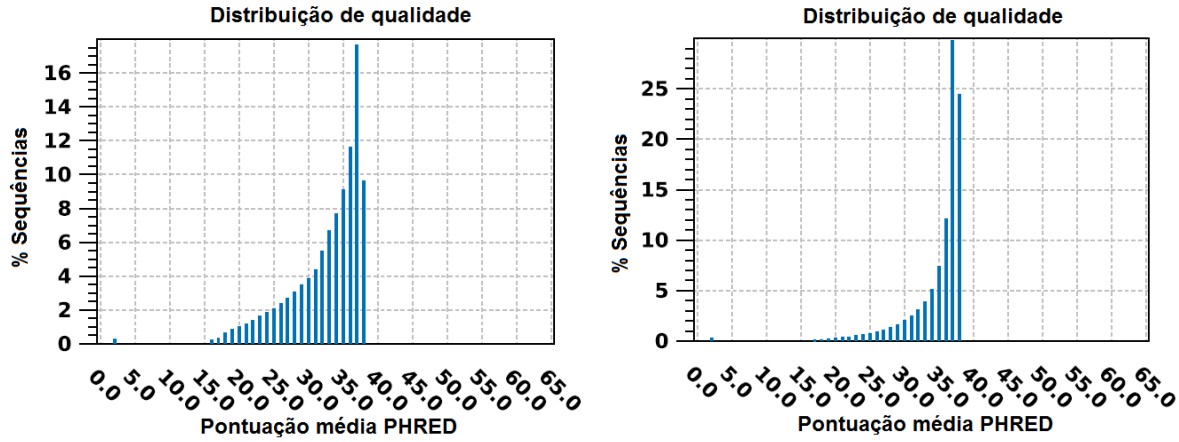

FIGURA 9 DISTRIBUIÇÃO DE QUALIDADE MÉDIA DAS BASES

As qualidades das leituras são calculadas através da média aritmética das qualidades das bases. O eixo X é a pontuação PHRED e o eixo Y o número de leituras normalizadas nas determinadas pontuações.

FONTE: O próprio autor através do programa CLC Genomics WorkBench

5.3 MONTAGEM DO GENOMA

5.3.1 Processos iniciais e montagem de *A. olearius* DQS4

Os processos iniciais de montagens foram realizados no programa CLC. Todas as etapas de atividades explicadas a seguir encontram-se no fluxograma das (FIGURAS 5 e 6) apresentadas anteriormente. Primeiramente foi realizada a retirada dos adaptadores utilizados no sequenciamento da biblioteca *Mate-Pair* através das ferramentas de cortes do CLC, *Trim Sequences*, como pode ser visualizado na etapa 1 da (FIGURA 5). Em seguida, foi realizada a montagem de cada biblioteca separadamente no montador CLC (Etapa 2), primerio sem a utilização da informação do tamanho de fragmento e em seguida com estas informações, assim foi possível verificar a diferença entre os resultados.

Assim, a melhor montagem da biblioteca *Mate-pair* no CLC foi a montagem utilizando as informações de pareamento, que resultadou em767 contigs e valor de N50 de 152.745 pares de bases. Nesta montagem foi realizada uma personalização de um dos parâmetros, ajustando o "*world size*" do CLC para o valor de 24. A montagem desta biblioteca sem as informações de pareamento resultou em uma montagem com resultados mais pobres, gerando 1.119 *contigs* e N50 de 4.148 pb.

Foram realizadas várias montagens com a biblioteca *Paired-end* nos montadores já citados neste trabalho, sendo as montagens no programa CLC, que novamente obtiveram os melhores resultados. Na montagem sem a utilização das informações de pareamento foram obtidos 580 *contigs* e N50 de 61.977pb e utilizando as informações de pareamento, o valor do N50 teve uma melhora significativa, montando 511 *contigs* e N50 de 241.091pb. Com o intuito de diminuir o número de *contigs* e aumentar o N50, foi gerada uma nova montagem com as informações de pareamento alterando o valor de "*world size*" para 24, conforme foi realizado com a montagem *Mate-pair* anteriormente. Esta montagem gerou os melhores resultados até então, montando 386 *contigs* e valor de N50 igual a 319.819 pb, sendo a montagem escolhida como a principal.

O programa CLC possui um módulo chamado de "*Microbial Genome Finish*", recomendado para finalizações de montagens, este módulo possui várias ferramentas para acertos manuais no genoma. Com o intuito de realizar uma melhoria final no genoma, foi utilizado este módulo em modo experimental. Foram utilizadas algumas ferramentas automáticas como "*Extend contigs*", o que realizou uma melhoria significativa na montagem, elevando um pouco mais o seu valor de

N50. Assim, a montagem final gerada no CLC ficou com 396 *contig*s e valor de N50 de 379.001pb sendo a melhor montagem.

Foram realizados alguns testes de montagem híbridos, utilizando-se as duas bibliotecas juntas na mesma montagem, porém não foram obtidos bons resultados. Das várias montagens nesta modalidade, a que obteve melhor resultado gerou 1.864 *contigs* e N50 de 38.233 pb, o que não foi um bom resultado comparado as demais montagens.

As montagens não se limitaram ao programa CLC, outros montadores foram utilizados a fim de gerar várias versões de montagens para serem analisadas e reutilizadas posteriormente, (Etapa 3 e 4). Foram realizadas várias montagens no programa Velvet (Zerbino, 2008), nele também foram testadas montagens com as duas bibliotecas, porém apenas a segunda biblioteca de leituras *Paired-end* gerou resultados razoáveis (Etapa 4), gerando 657 contigs conforme mostra a (TABELA 3). O valor de N50 não foi registrado nestas montagens devido a não obterem bons resultados comparado ao programa CLC.

Outro programa utilizado foi o SOAP (Luo *et al*., 2012). Neste programa a primeira biblioteca de leituras *Mate-pair* obteve melhores resultados de montagem do que a *Paired-end*. Na (TABELA 3), id 3, pode ser verificado que a melhor montagem com a biblioteca *Mate-pair* foi obtida pelo programa SOAP, (Etapa 3).

Por fim, foi utilizado o montador Newbler (Roche), mas este não gerou bons resultados, sendo que a única montagem significativa foi a com id de número 9, conforme (TABELA 3).

TABELA 3 MONTAGENS REALIZADAS

ID	Programas montadores	Contigs	N50
	Biblioteca Mate-pair		
1	CLC Genomic	1.119	4.148
2	CLC Genomic	767	152.745
3	SOAP	637	20.045
4	Velvet	3.946	4.143
	Biblioteca Paired-end		
5	CLC Genomic	580	61.977
6	CLC Genomic	511	241.091
7	CLC Genomic	386	319.819
8	CLC Genomic Modulo Finish	396	379.001
9	Newbler	413	66.327
10	SOAP	1.528	13.833
11	Velvet	914	*
12	Velvet	657	*
	Montagens híbridas		
13	CLC Genomic	1.864	38.233
14	CLC Genomic	9.536	24.143
15	SOAP	3.373	5.865

FONTE: O próprio autor.

Para a geração dos *scaffolds*, com o objetivo de reduzir o numero de *contigs*, juntá-los e gerar uma molécula única, foi utilizada o montador Newbler novamente, reunindo três tipos de dados, conforme (Etapa 7). Primeiramente foram escolhidas as duas melhores montagens de numero id 7 e id 8 (TABELA 3). Estas montagens foram escolhidas devido aos seus melhores resultados baseados na quantidade de contigs gerados e valores do indicador N50. Juntamente com estas montagens foram inseridas todas as leituras das duas bibliotecas originais, e por fim foram selecionadas as duas montagens do montador Velvet, com numero id 11 e 12 (TABELA 3), provenientes da Etapa 4. As montagens do Velvet foram escolhidas devido a este programa utilizar uma estratégia de montagem diferenciada do programa CLC, gerando *contigs* alternativos com diferentes tamanhos e juntar estes dois tipos de dados em um montador, poderia aumentar a probabilidade de extender os *contigs* já existentes.

Conforme CHAISSON e PEVZNER (2008), o montador Newbler funciona melhor com leituras de até 500 pares de bases. Baseado nesta informação, foi utilizado a estratégia de transformar todos os *contigs* das quatro montagens anteriores mencionadas em *contigs* até 500 pb, realizados na (Etapa 5). Para realizar este corte e gerar novos *contigs*, foi utilizado um script de quebra, Rodrigo Cardoso (não publicado), onde contigs com tamanhos menores de 500 pb permaneceram com tamanho original e *contigs* maiores foram quebrados, mantendo uma sobreposição de 50 bases para posteriormente permitir a remontagem dos *contigs* originais. Assim, os dados de entrada para a montagem no programa Newbler ficaram como se fossem leituras inicias novamente, porém, com a grande maioria delas com o dobro do tamanho.

Estas leituras, com tamanhos de até 500 pares de bases foram submetidos a um script de padronização de cabeçalho para serem utilizados pelo programa Newbler, Script pré-Newbler, (Etapa 6). As leituras originais do sequenciamento também foram submetidas ao tratamento de seus cabeçalhos através do mesmo script devido às informações de pareamento das bibliotecas.

Estes dados foram submetidos e remontados pelo montador Newbler, (Etapa 7), que identificou as sobreposições entre as leituras, construindo alinhamentos múltiplos de sobreposição. Foram geradas várias montagens de testes e a montagem que apresentou melhores resultados foi classificada como a montagem final (TABELA 4).

TABELA 4 DADOS DA MONTAGEM FINAL

Contigs	**Scaffolds**	**N50**
175	18	656.561

FONTE: O próprio autor.

5.3.2 Fechamento do genoma de *A. olearius* DQS4

O arquivo fasta contendo os 18 scaffolds com 137 gaps foi submetido ao programa FGAP (PIRO et al, 2014) juntamente com um conjunto de dados contendo montagens alternativas produzidas previamente na parte inicial do processo, (Etapa 9). Todos os 137 gaps internos foram fechados gerando 18 contigs. As montagens utilizadas foram as de numero id 2, 3, 7, 8, 11 e 12 (TABELA 3). Estas montagens

passaram pelo ”Script gera DATASET de retirada de N”, (Etapa 8), que faz a remoção das bases ambíguas, com “N”, das montagens utilizadas como base de dados para o fechamento dos gaps na montagem principal.

Os 18 *contigs* resultantes foram ordenados com o genoma de referência utilizando o programa Mummer/Prommer (KURTZ et al, 2004) para a geração de um dotplot de comparação, o qual demonstrou uma alta sintenia. A tabela guia dos *scaffolds* gerada pelo Mummer apresentou alguns *scaffolds* que estavam em sentido invertido. Para a geração dos dotplots, este redirecionamento é realizado automaticamente apenas para fins de visualização, ele não faz o redirecionamento nos arquivos, os quais necessitaram ser redirecionados manualmente. Também foi verificada a necessidade de posicionar o início do genoma em seu local correto. Conforme convenção, o gene da *dnaA* deve ficar no início do genoma. Segundo o alinhamento produzido pelo Mummer, o gene da *dnaA* estava fora de posição em relação ao genoma de referência (FIGURA 10).

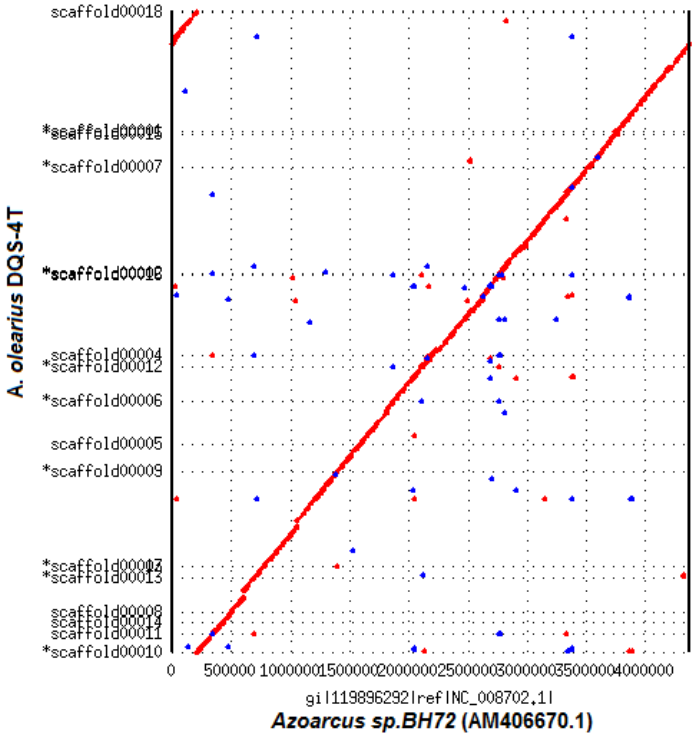

FIGURA 10 ALINHAMENTO DOS CONTIGS DE A. OLEARIUS DQS4 AO GENOMA DE REFERENCIA DE AZOARCUS BH72
FONTE: O próprio autor através dos programas Mummer / Prommer

Após o FGAP, foi realizado o ordenamento dos *contigs* com o auxílio de um guia gerado pelo programa Mummer/Prommer através do comando “*showtiling*”,

como pode ser observado na (Figura 11). Os contigs que apresentam o sinal de mais, “+”, estão na mesma direção do genoma de referência. Os *contigs* com sinal de menos, “-“, foram colocados manualmente em ordem e sentidos corretos, complemento reverso, em relação ao genoma de referência, originando um único *scaffold*. Os contigs 16, 17 e 18 não foram alinhados no genoma de referência e ficaram fora da geração da molécula única. Foi utilizado a feramenta *Reverse Comlement*, para realizar o complemento reverso dos contigs no seu ordenamento. Disponível na internet através do link <http://www.bioinformatics.org/sms/rev_comp.html>.

```
>gi|119896292|ref|NC_008702.1| 4376040 bases
209741  336614  5941    126874  99.93   96.83   -       scaffold00010
342556  427049  5594    84494   92.72   94.48   +       scaffold00011
432644  498349  -40566  65706   100.00  97.66   +       scaffold00014
457784  690590  207     232807  82.28   97.38   +       scaffold00008
690798  770560  -27877  79763   100.00  97.27   -       scaffold00013
742684  1393262 3891    650579  92.44   97.00   -       scaffold00003
1397154 1585830 1203    188677  96.70   97.48   -       scaffold00009
1587034 1882334 -13119  295301  95.17   97.55   +       scaffold00005
1869216 2111936 -1177   242721  93.32   96.16   -       scaffold00006
2110760 2191988 26505   81229   100.00  94.40   -       scaffold00012
2218494 2771369 2010    552876  93.71   95.91   +       scaffold00004
2773380 3503104 14039   729725  96.20   95.67   -       scaffold00002
3517144 3752877 -9416   235734  93.64   95.70   -       scaffold00007
3743462 3759491 11321   16030   100.00  99.32   -       scaffold00015
3770813 4596704 -220664 825892  72.58   96.80   -       scaffold00001
```

FIGURA 11 GUIA GERADO PELO PROGRAMA MUMMER / PROMMER, ATRAVÉS DO COMANDO “SHOWTILING”
FONTE: O próprio autor através do programa Mummer / Prommer

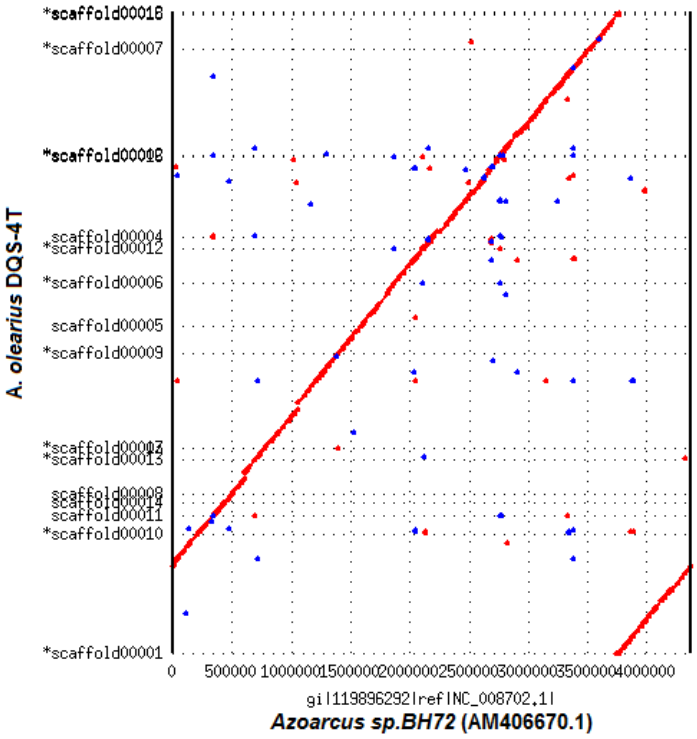

FIGURA 12 DOTPLOT APÓS COMPLEMENTO REVERSO NOS SCAFFOLDS INVERTIDOS.
Fonte: O próprio autor através do programa Mummer / Prommer

Como pode ser observado na (FIGURA 12), existe um fragmento do genoma que não está alinhado corretamente. Parte deste *contig* contem o gene *dnaA*, e foi necessário coloca-lo no início do genoma. Para reordenação do gene *dnaA* os contigs foram anotados no programa SILA e a posição do gene foi identificada através do programa Artemis. O fragmento do genoma contendo o gene *dnaA* foi removido e realocado no inicio do genoma criando um novo *contig,* na sequência do *scaffold* original. O genoma reordenado foi novamente alinhado ao genoma de referência de *Azoarcus* BH72 mostrando o posicionamento correto dos genes na (FIGURA 13).

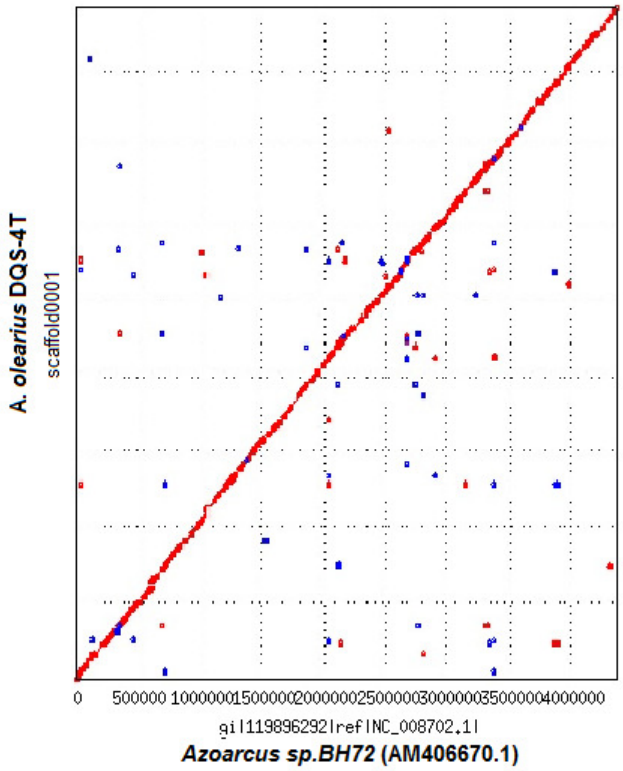

FIGURA 13 ALINHAMENTO DO GENOMA DE *A. olearius* DQS4 REORDENADO COM GENOMA DE REFERÊNCIA DE Azoarcus BH72

FONTE: O próprio autor através do programa Mummer / Prommer

Os *contigs* alinhados com o genoma de referência foram unidos em uma só sequência e separados por “N” a fim de indicar regiões de gaps.

5.3.3 Fechamento de Gaps Manual

Após várias análises, descobriu-se que o contig 16 representava uma região repetiviva no genoma do *A. olearius* DQS4. A comparação com o banco de dados Genbank usando programa BLASTn mostrou que esta sequência continha aproximadamente 87 ocorrências no genoma de referência. Comparando somente as pontas do *contig*, foi localizada a ocorrência de dois genes do genoma de referência de *Azoarcus* BH72, o gene *etfA3* e *azo2503*. No *draft* do genoma de A. *olearius* DQS4 esses genes estavam separados por uma região de gap no interior do gene *etfA3 e* entre este *gap* e o outro gene, *azo2503,* havia uma sequência onde a comparação com o banco de dados não resultou em nenhuma similariedade

conhecida. Esta sequência foi excluída até a região do *gap* do gene *etfA3*, permanecendo ainda parte do mesmo, até o gene *azo2503.* Nas extremidades da sequência removida, foram adicionados “Y” e em seu interior foi inserida a sequência do *contig* 16. Após estas modificações, o arquivo com a molécula única foi submetido ao programa FGAP novamente a fim de homologar e fechar os *gaps* com Y, o qual realizou com sucesso a tarefa (FIGURA 14).

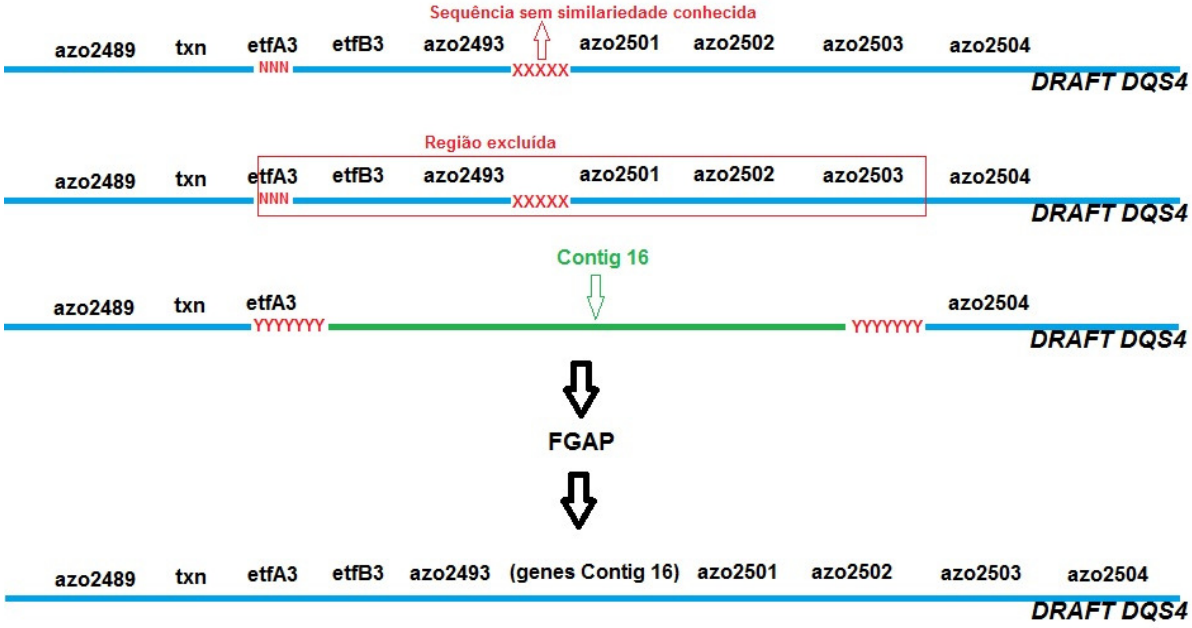

FIGURA 14 INSERÇÃO DO SCAFFOLD 16 - NUMERAÇÃO RELATIVA À AZOARCUS BH72.
FONTE: O próprio autor.

O *contig* 17 não entrou na montagem automática do genoma por ser uma região de repetição. Através da submissão do mesmo ao programa RNAmmer, foram localizados os 3 rRNAs ribossomais 16S, 23S, 5S (FIGURA 15).

```
##gff-version2
##source-version RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35)
##date 2014-08-19
##Type DNA
# seqname           source                      feature     start      end   score   +/-  frame  attribute
# ------------------------------------------------------------------------------------------------------------
Sequence        RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    5214    5325    88.4    +       .       5s_rRNA
Sequence        RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    2223    5105    3542.1  +       .       23s_rRNA
Sequence        RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    176     1699    1950.7  +       .       16s_rRNA
# ------------------------------------------------------------------------------------------------------------
```

FIGURA 15 OPERONS RIBOSSOMAIS *A. olearius*. DQS4
FONTE: O próprio autor através do programa RNAmmer

Também foi submetido ao programa RNAmmer o genoma do *Azoarcus* BH72, a fim de se verificar a quantidade de operons ribossomais que o genoma de

referência possuía. Conforme (FIGURA 16), foi revelado que o mesmo possuía 4 regiões com operons ribossomais.

```
##gff-version2
##source-version RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35)
##date 2014-06-16
##Type DNA
# seqname          source                    feature     start      end   score   +/-  frame  attribute
# ------------------------------------------------------------------------------------------------------
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    209850  209961  88.4    +       .       5s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    775772  775883  88.4    +       .       5s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    1396954 1397065 88.4    +       .       5s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    3511782 3511893 88.4    -       .       5s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    206859  209741  3542.1  +       .       23s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    772781  775663  3542.1  +       .       23s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    1393963 1396845 3542.1  +       .       23s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    3512002 3514884 3542.1  -       .       23s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    204812  206335  1950.7  +       .       16s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    770734  772257  1950.7  +       .       16s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    1391916 1393439 1950.7  +       .       16s_rRNA
gi_119896292_ref_NC_008702.1_   RNAmmer-1.2 (IRIX64 organism 6.5 07202013 IP35) rRNA    3515408 3516931 1950.7  -       .       16s_rRNA
# ------------------------------------------------------------------------------------------------------
```

FIGURA 16 OPERONS RIBOSSOMAIS DO GENOMA DO *Azoarcus* BH72
Fonte: O próprio autor através do programa RNAmmer

Para comprovar a número de repetições dos rRNAs, foi verificado no programa CLC a cobertura dos *contigs* na melhor montagem realizada neste montador, onde o *contig* 19 desta montagem corresponde ao *contig* 17 da montagem final. Ele continha os rRNAs e foi calculado em aproximadamente 4 vezes a cobertura do genoma. 691,94 / 161 = 4,29 conforme esperado em relação ao seu genoma de referência (FIGURA 17).

Name	Consensus length	Total read count	Average coverage
Azoarcus_S3_L001_R1_001.pe (paired) contig 1 mapping	825936	909841	180,56
Azoarcus_S3_L001_R1_001.pe (paired) contig 10 mapping	573753	595322	169,32
Azoarcus_S3_L001_R1_001.pe (paired) contig 7 mapping	490156	473132	157,89
Azoarcus_S3_L001_R1_001.pe (paired) contig 14 mapping	379001	447477	191,68
Azoarcus_S3_L001_R1_001.pe (paired) contig 4 mapping	332743	302350	148,59
Azoarcus_S3_L001_R1_001.pe (paired) contig 11 mapping	294930	274209	152,62
Azoarcus_S3_L001_R1_001.pe (paired) contig 8 mapping	241367	212514	143,55
Azoarcus_S3_L001_R1_001.pe (paired) contig 9 mapping	241093	241268	162,25
Azoarcus_S3_L001_R1_001.pe (paired) contig 3 mapping	235794	254948	175,57
Azoarcus_S3_L001_R1_001.pe (paired) contig 6 mapping	221182	200660	149,65
Azoarcus_S3_L001_R1_001.pe (paired) contig 2 mapping	188723	188284	162,26
Azoarcus_S3_L001_R1_001.pe (paired) contig 20 mapping	128819	136057	173,63
Azoarcus_S3_L001_R1_001.pe (paired) contig 13 mapping	85488	97787	187,15
Azoarcus_S3_L001_R1_001.pe (paired) contig 12 mapping	77280	71104	151,04
Azoarcus_S3_L001_R1_001.pe (paired) contig 17 mapping	47842	42158	142,09
Azoarcus_S3_L001_R1_001.pe (paired) contig 24 mapping	18254	16287	144,44
Azoarcus_S3_L001_R1_001.pe (paired) contig 16 mapping	16076	19305	198,56
Azoarcus_S3_L001_R1_001.pe (paired) contig 23 mapping	15736	13967	142,81
Azoarcus_S3_L001_R1_001.pe (paired) contig 18 mapping	11362	10845	155,21
Azoarcus_S3_L001_R1_001.pe (paired) contig 19 mapping	5376	26119	691,94
Azoarcus_S3_L001_R1_001.pe (paired) contig 36 mapping	1522	1292	131,90
Azoarcus_S3_L001_R1_001.pe (paired) contig 29 mapping	1250	1422	179,38

FIGURA 17 COBERTURA DOS CONTIGS NA MONTAGEM DO GENOMA DE *A. olearius* DQS4
Fonte: O próprio autor através do programa CLC Genomics Workbench

Confrontando as regiões do genoma de referência com o genoma do *A. olearius* DQS4, foram encontradas 4 regiões de gaps. O mesmo processo realizado na inserção do contig 16 foi executado nestas 4 regiões. Após a inserção das sequências faltantes, o mesmo foi submetido ao FGAP, o qual fechou os *gaps* com sucesso.

O *contig* 18 foi considerado material genético de outros organismos. Este contig não foi utilizado na montagem e foi descartado.

5.3.4 Mapeamento e dados finais do genoma de *A. olearius* DQS4

Após a finalização da etapa de fechamento de *gaps*, o genoma do *A. olearius* DQS4 foi submetido ao programa CLC para ser realizado um mapeamento das leituras originais do sequenciamento conforme (TABELA 5). A análise do mapeamento das leituras pareadas mostrou que todo o genoma montado de *A. olearius* DQS4 estava coberto e que apesar de variável, não havia regiões de cobertura zero, o que seria um indicativo de erro de montagem. Foram mapeados 99,79% das leituras no genoma montado. O genoma de *A. olearius* DQS4 apresenta 4.451.750 pb com um conteúdo GC% de 67,83% conforme (TABELA 6), onde pode ser observados os demais dados finais da montagem.

TABELA 5 MAPEAMENTO DAS LEITURAS COM O GENOMA DE A. OLEARIUS DQS4 NO PROGRAMA CLC GENOMICS COM A FERRAMENTA *MAPPING TOOL*

	Contagem	**Porcentagem das leituras**	**Tamanho médio**	**Numero de bases**	**Porcentagem de bases**
Referências	1	-	4,451,751,00	4,451,751	-
Leituras Mapeadas	4,008,977	99.79%	134,41	538,837,711	99.78%
Leituras não mapeadas	8,24	0.21%	145,47	1,198,676	0.22%
Leituras pareadas	3,267,960	81.35%	139,56	403,793,612	74.77%
Leituras pareadas quebradas	646,604	16.10%	192,93	124,750,630	23.10%
Toral de leituras	4,017,217	100.00%	134,43	540,036,387	100.00%

FONTE: O próprio autor.

TABELA 6 DADOS DO GENOMA DE A. *OLEARIUS* DQS4

Características	**DQS-4T**
Acesso Genbank	CP016210.1
	Chromosome
Tamanho (pb)	4,451,750
GC (%)	67.83
# ORFs	4,067
# rRNA operons	4
# tRNA	55

FONTE: O próprio autor.

5.4 COMPARAÇÃO GENÔMICA

A sequência de nucleotídeos do genoma de *A. olearius* DQS4 foi comparada com as sequências genômicas de *Azoarcus* BH72, *Azoarcus* KH32C, *Azoarcus* EbN1 e *Azoarcus toluclasticus* ATCC700655 conforme pode ser visualizado na (TABELA 7). Também foi utilizando o calculo de identidade média (FIGURA 18) e sintenia (FIGURA 19). Os resultados das comparações revelaram que *A. olearius* DQS4 e *Azoarucs* BH72 apresentam 98,98% de identidade média entre seus genomas. A similaridade com *Azoarcus* KH32C, *Azoarcus* EbN1 e *A. toluclastico* ficou em 82,64%, 82,45% e 82,79%, respectivamente. Segundo Goris e colaboradores (2007) um valor maior que 95% de identidade na comparação são considerados organismos da mesma espécie. A análise de sintenia utilizando Mummer também indicou maior correspondência com o genoma de *Azoarcus* BH72. Esses dados sugerem que *A. olearius* DQS4 e *Azoarcus* BH72 pertencem a uma mesma espécie.

TABELA 7 COMPARAÇÃO GENÔMICA

	Estirpes							
Características	**DQS-4T**	**BH72**	**KH32C**		**EBN1**			**ATCC700605**
Acesso Genbank	CP016210.1	AM406670	AP012304		CR555306			ARJX01000000
	Chromosome	Chromosome	Chromosome	Plasmid	Chromosome	Plasmid 1	Plasmid 2	Chromosome
Tamanho (pb)	4,451,750	4,376,040	5,081,166	737,589	4,296,230	207,355	223,67	5,924,508
GC (%)	67.83	67.92	65.10	64.5	65.12	57.63	63.11	66.00
# ORFs	4,067	4,036	4,531	657	4,133	274	196	5,076
# rRNA operons	4	4	5	-	4	-	-	3
# tRNA	55	56	64	-	58	-	-	50

FONTE: National Center for Biotechnology Information, U.S. National Library of Medicine. 2013.

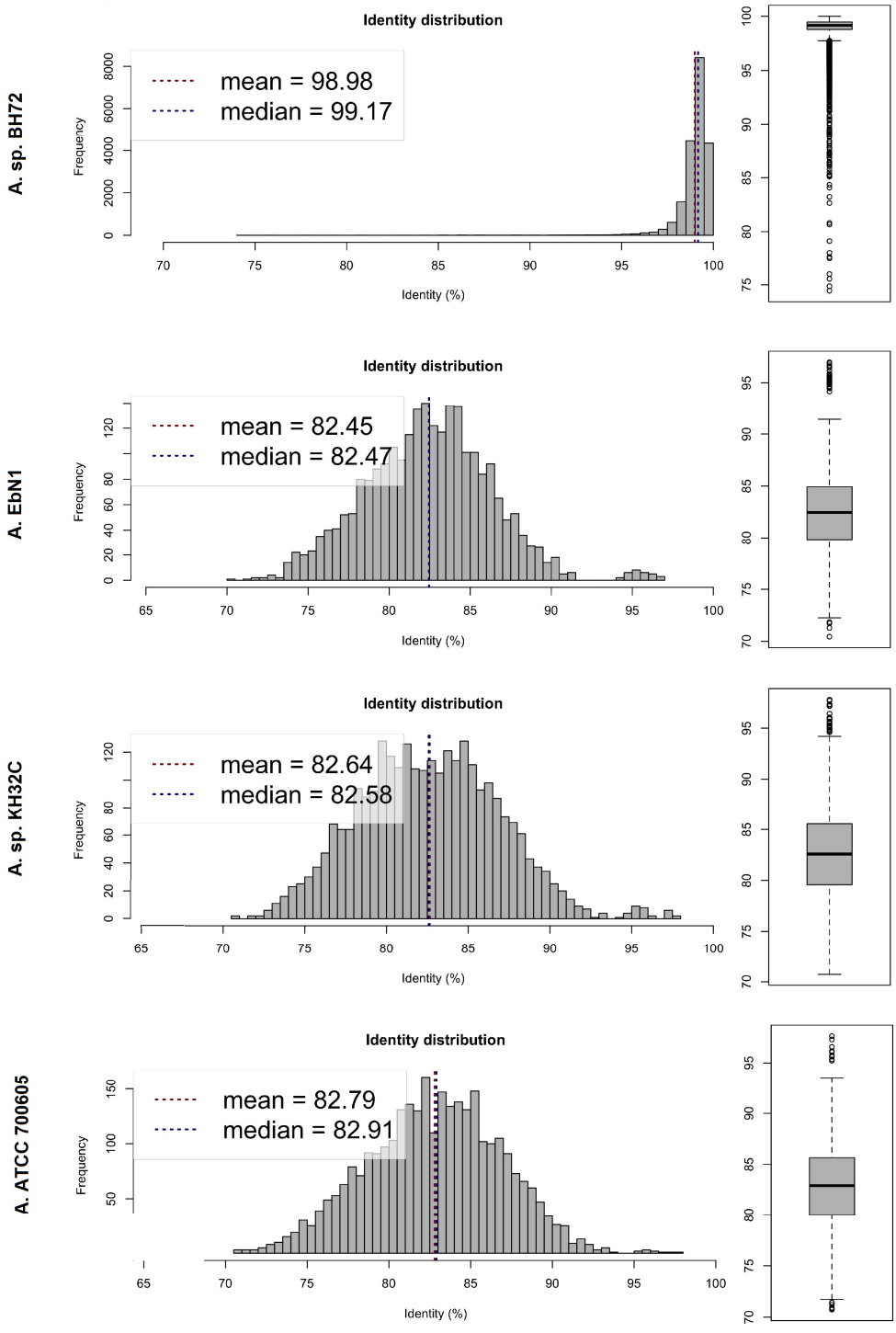

FIGURA 18 CÁLCULO DE IDENTIDADE MEDIA ENTRE OS GENOMAS DE *A. olearius* DQS4 E OUTROS GENOMAS DE BACTERIAS DO GÊNERO *Azoarcus*

FONTE: O próprio autor através do programa ANI

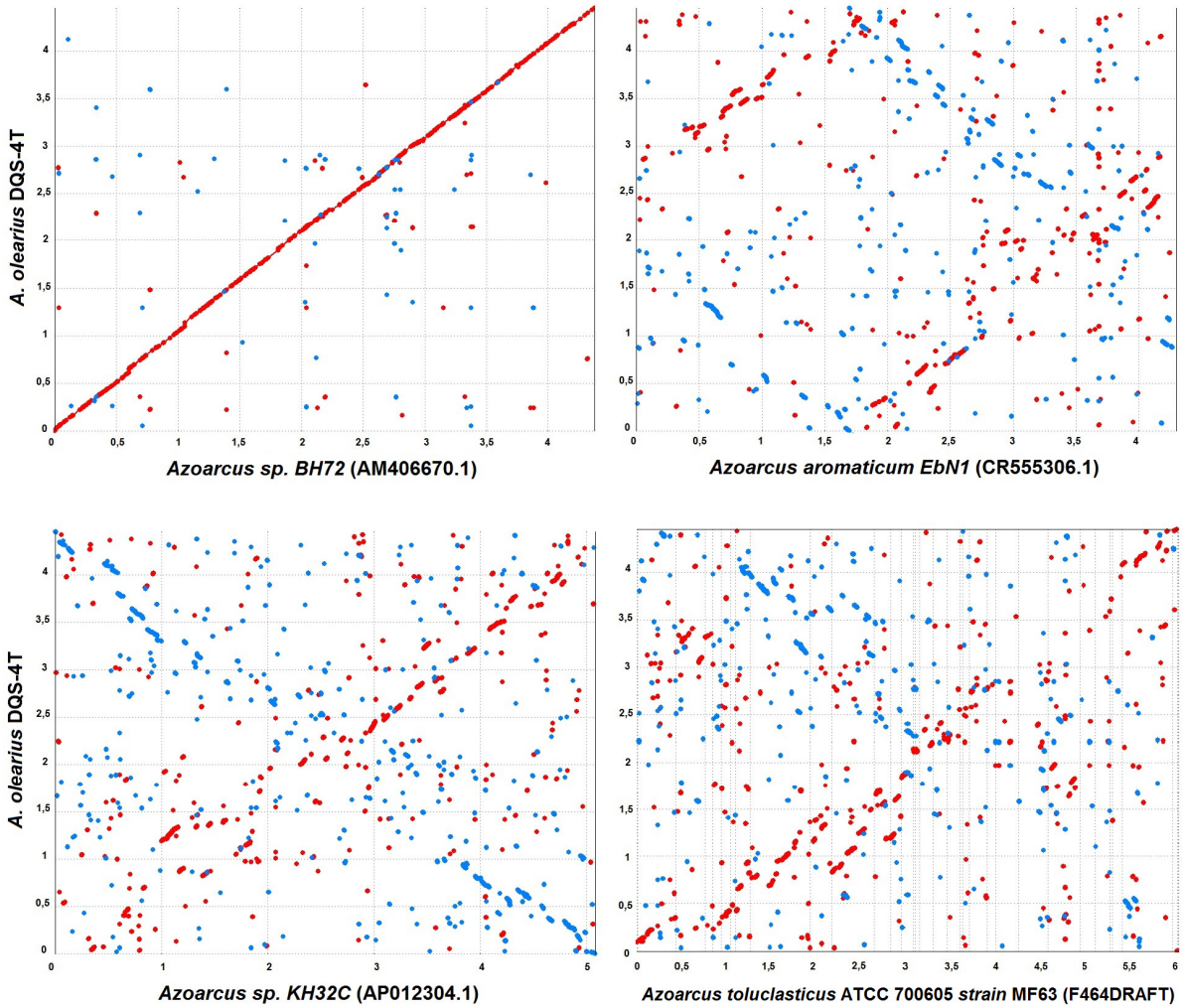

FIGURA 19 ANALISE DE SINTENIA ENTRE GENOMAS DE *A. olearius* DQS4, *Azoarcus* BH72, *Azoarcus* EBN1 E *A. toluclastico* ATCC700655.

FONTE: O próprio autor através do programa Mummer / Prommer

5.4.1 Anotação do genoma de *A. olearius* DQS4

O genoma de *A. olearius* DQS4 foi submetido ao programa SILA (VIALLE *et al.* 2013), que realizou a anotação automática identificando as prováveis regiões codificadoras de proteínas (CD). Também foi realizada a anotação automática com a plataforma RAST (AZIZ *et al*. 2008) a fim de se utilizar as duas anotações para definir a anotação final. As regiões anotadas foram revisadas manualmente para ajustes na posição do códon de início de tradução e validação dos genes anotados

com a utilização do *software* Artemis. Genes codificadores de rRNA e tRNA foram anotados pela plataforma RAST. Através dessas análises foram identificadas 4.067 regiões codificadoras, 55 tRNAs além dos 4 operons rRNA contendo os genes 16S, 23S e 5S. O gráfico gerado pelo RAST mostra a diversidade de genes anotados e classificados em subsistemas para grupos funcionais encontrando 3230 genes (FIGURA 20).

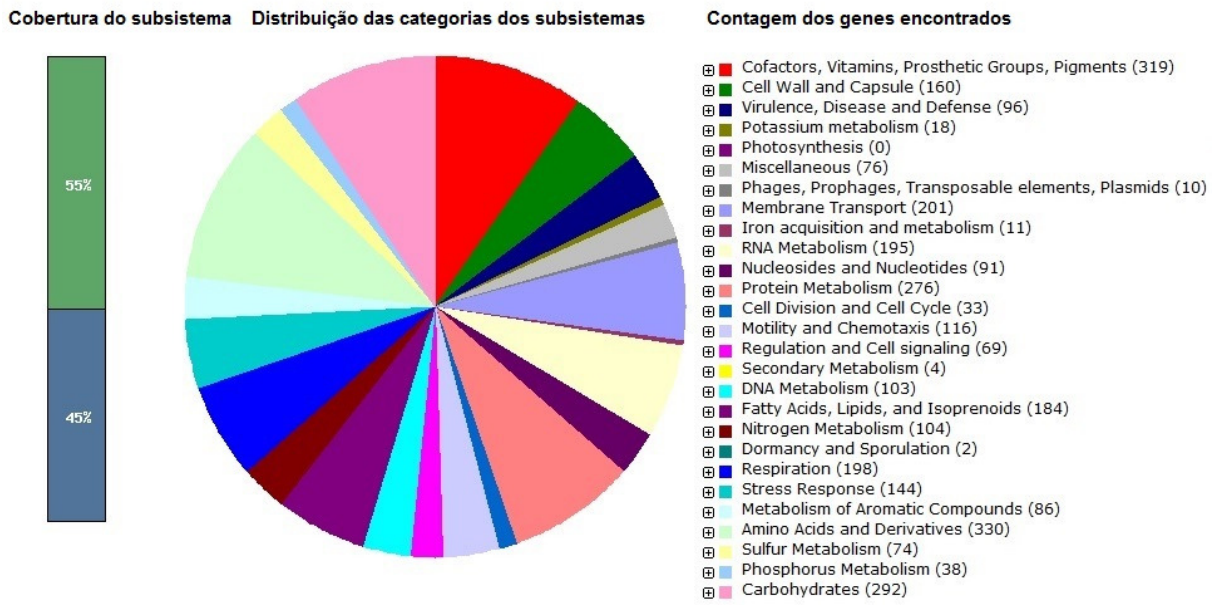

FIGURA 20 DIVERSIDADE FUNCIONAL DOS GENES ANOTADOS NO GENOMA DE *A. olearius* DQS4.
FONTE: O próprio autor através do programa RAST

5.5 COMPARAÇÃO COM GENOMA DE REFERÊNCIA

O genoma de referência de *Azoarcus* BH72 também foi reanotado no servidor RAST, assim foi possível realizar uma comparação entre os dois genomas. A (FIGURA 21) mostra o número de genes anotados em ambos os genomas, assim como os subgrupos classificados. É possível visualizar que ambos os genomas são muito semelhantes existindo poucas variações em seus subgrupos.

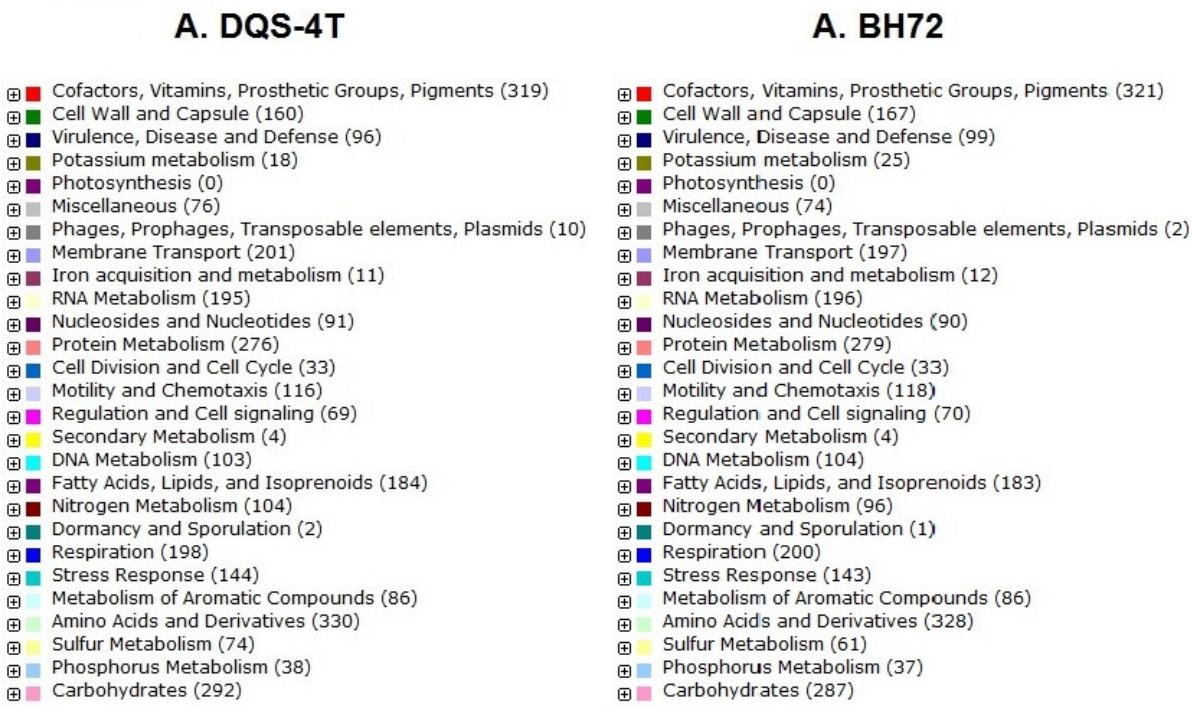

FIGURA 21 COMPARAÇÃO ENTRE OS SUBGRUPOS IDENTIFICADOS

FONTE: O próprio autor através do programa RAST

Os genes de maior interesse anotados são os subgrupos de Metabolismo de Nitrogênio e Metabolismo de Componentes aromáticos, que podem ser visualizados na comparação mais detalhada na (FIGURA 22).

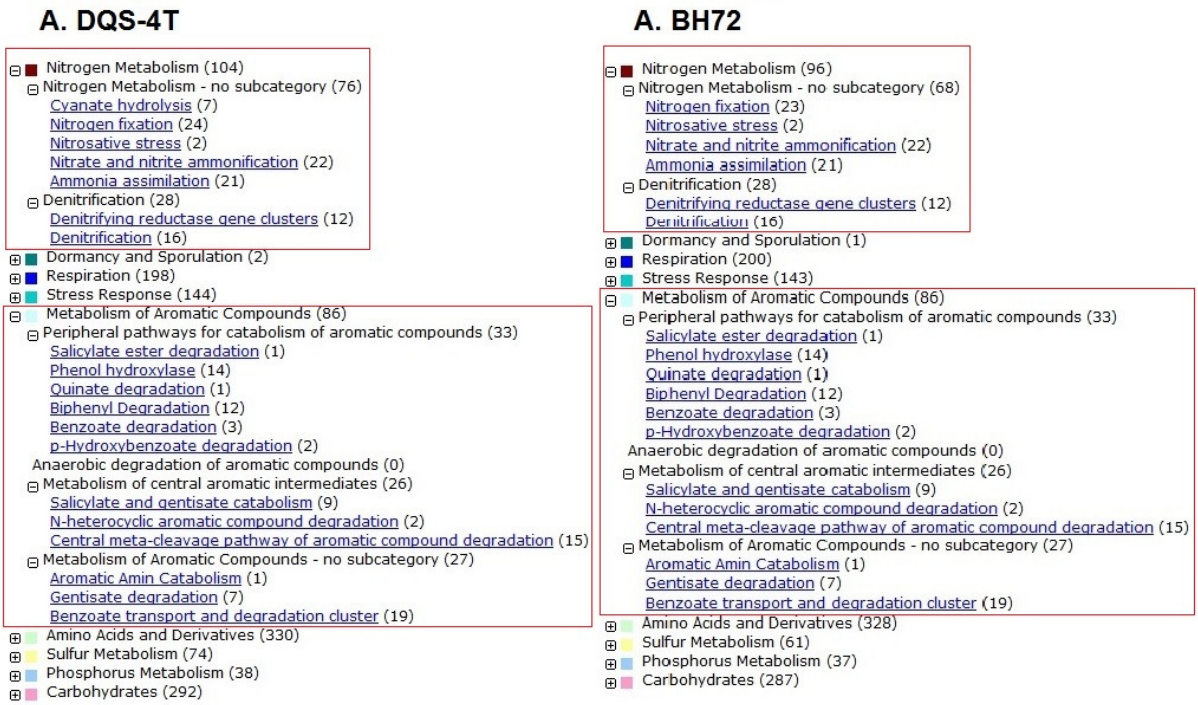

FIGURA 22 COMPARAÇÃO ENTRE OS SUBGRUPOS DE METABOLISMO DE NITROGÊNIO E METABOLISMO DE COMPONENTES AROMÁTICOS

FONTE: O próprio autor através do programa RAST

O genoma do *A. olearius* DQS4 possui mais genes anotados para o subgrupo de Metabolismo de Nitrogênio do que o *Azoarcus* BH72, são 104 genes anotados comparados a 96 genes na estirpe BH72. A diferença entre o metabolismo de nitrogênio dos dois genomas deve-se ao fato do A. *olearius* DQS4 possuir genes que codificam proteínas relacionadas à hidrolise de cianeto (7 genes) e também pela existência de um gene a mais no subgrupo de fixação de nitrogênio correspondente a 4Fe-4S ferredoxina, associado a nitrogenase.

Em relação aos genes de interesse relacionados ao Metabolismo de componentes aromáticos, ambos os genomas possuem a mesma quantidade de genes anotados dentro de cada subgrupo.

5.6 REGIÕES DE INSERÇÃO DO A. DQS4 NO GENOMA DE REFERÊNCIA.

Com a utilização do programa ACT, foi possível identificar 10 regiões de divergência entre os genomas do *A. olearius* DQS4 e o genoma de referência

Azoarcus BH72 (FIGURA 23). Foram identificadas 7 grandes regiões presentes na estirpe DQS4 e ausentes na estirpe BH72 que se destacaram nas análises. As mesmas foram minunciosamente analisadas com o programa ACT juntamente com o programa BLASTp. As proteínas codificadas em cada região foram comparadas com o banco de dados através do programa BLASTp a fim de se identificar os genes envolvidos e seus domínios conservados. As regiões 5, 6 e 9 são regiões presentes na estirpe BH72 e ausentes no DQS4. As três maiores regiões de inserção do A. DQS4 foram às três primeiras regiões analisadas.

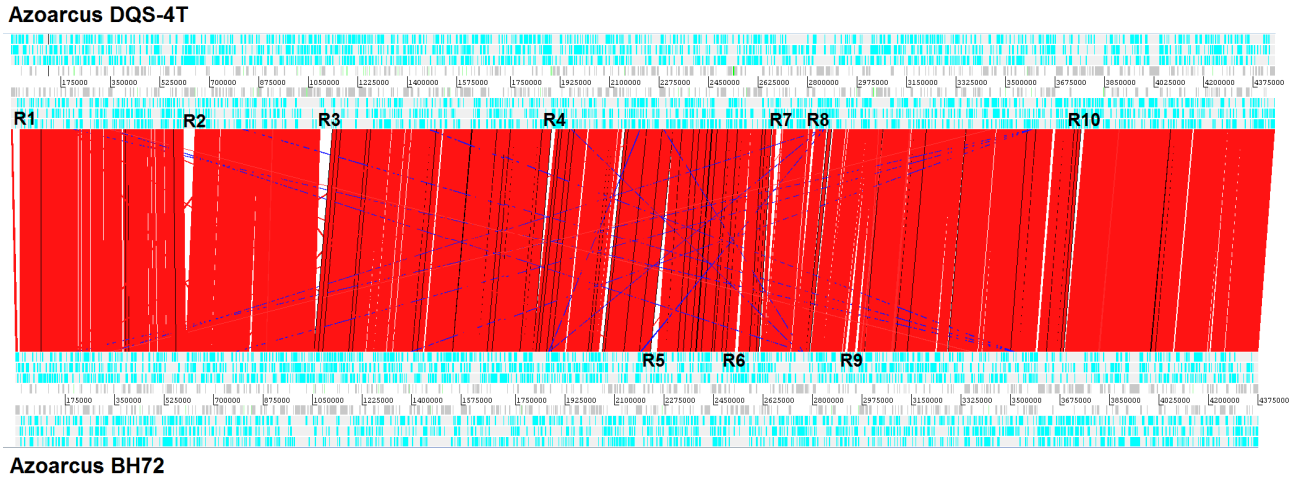

FIGURA 23 COMPARAÇÃO GENÔMICA DAS REGIÕES DE INSERÇÃO ENTRE AS ESTIRPES DQS4 E BH72
FONTE: O próprio autor através do programa ACT

5.6.1 Região 1

A primeria região adicional no genoma de *A. olearius* DQS4 (FIGURA 24) está localizada entre as posições 5.312 e 31.369, possui 26.059 pb e codifica 20 proteínas (TABELA 8). Seu conteúdo GC é de 59%.

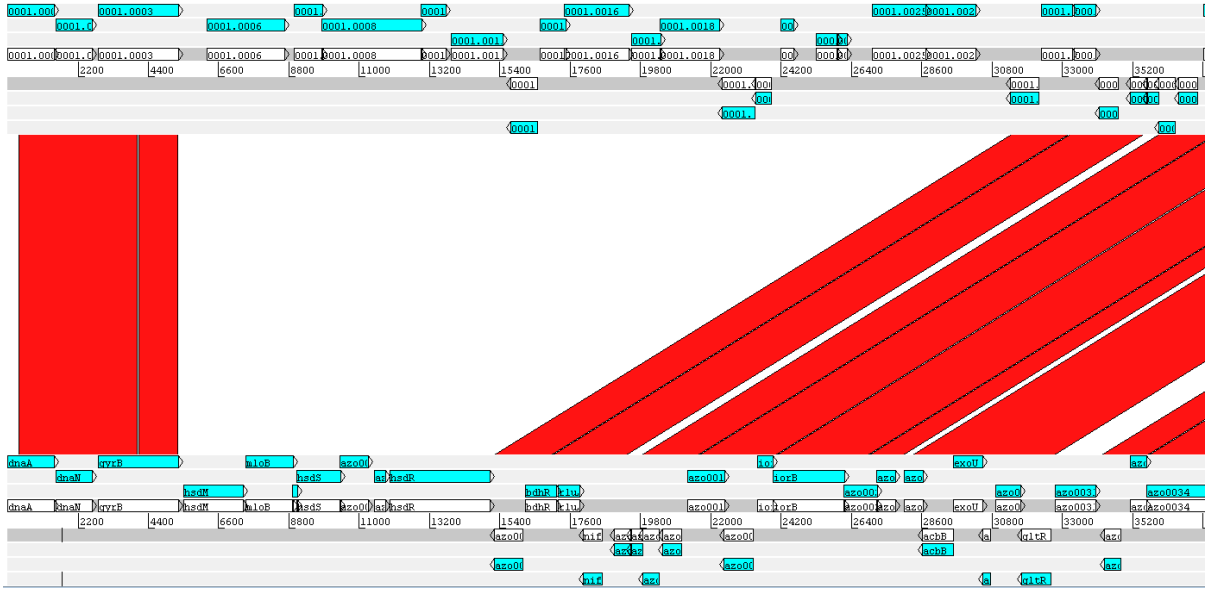

FIGURA 24 REGIÃO DE DIVERGÊNCIA 1
FONTE: O próprio autor através do programa ACT

TABELA 8 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 1

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.0006	Type I restriction-modification enzyme subunit M [Candidatus Competibacter denitrificans Run_A_D11]		91
0001.0007	Type I restriction-modification system, specificity subunit S [Thiorhodococcus sp. AK35]		73
0001.0008	Type I restriction-modification system, restriction subunit R [Burkholderia sp. AU4i]		91
0001.0009	putative metal-dependent hydrolase [Burkholderia sp. AU4i]		88
0001.0011	hypothetical protein [Pseudomonas stutzeri]	RES domain protain	71
0001.0013	hypothetical protein [Thauera sp. MZ1T]	**Nenhum domínio conservado foi encontrado**	63
0001.0015	hypothetical protein Tmz1t_0015 [Thauera sp. MZ1T]	TnsA endonuclease N terminal	64
0001.0016	hypothetical protein Tmz1t_0016 [Thauera sp. MZ1T]	Winged helix-turn helix	62
0001.0017	hypothetical protein [Thauera sp. MZ1T]	AAA domain; ATPases associated with a variety of cellular activities	67
0001.0018	hypothetical protein Tmz1t_0018 [Thauera sp. MZ1T]	TniQ along with TniA and B	69
0001.0019	hypothetical protein AHML_07055 [Aeromonas hydrophila ML09-119]	PLDc_BfiI_DEXD_like Catalytic domain of type II restriction endonucleases BfiI and NgoFVII	36
0001.0020	endonuclease [Lysobacter defluvii IMMIB APB-9 = DSM 18482]		43
0001.0021	hypothetical protein [Thauera sp. MZ1T]	**Nenhum domínio conservado foi encontrado**	42
0001.0022	general secretion pathway protein L [Hyphomonas jannaschiana VP2]	**Nenhum domínio conservado foi encontrado**	56
0001.0023	hypothetical protein Tmz1t_0021 [Thauera sp. MZ1T]	alpha/beta hydrolase fold	57
0001.0024	hypothetical protein Tmz1t_0023 [Thauera sp. MZ1T]	**Nenhum domínio conservado foi encontrado**	68
0001.0025	restriction endonuclease [Pseudomonas sp. URHB0015]	AAA domain (dynein-related subfamily);	47

FONTE: O próprio autor

5.6.2 Região 2

A segunda região adicional no genoma de *A. olearius* DQS4 (FIGURA 25) esta localizada entre as posições 608.493 e 650.814, possui 42.322 pb, codificando 59 proteínas (TABELA 9) e com conteúdo GC de 67%.

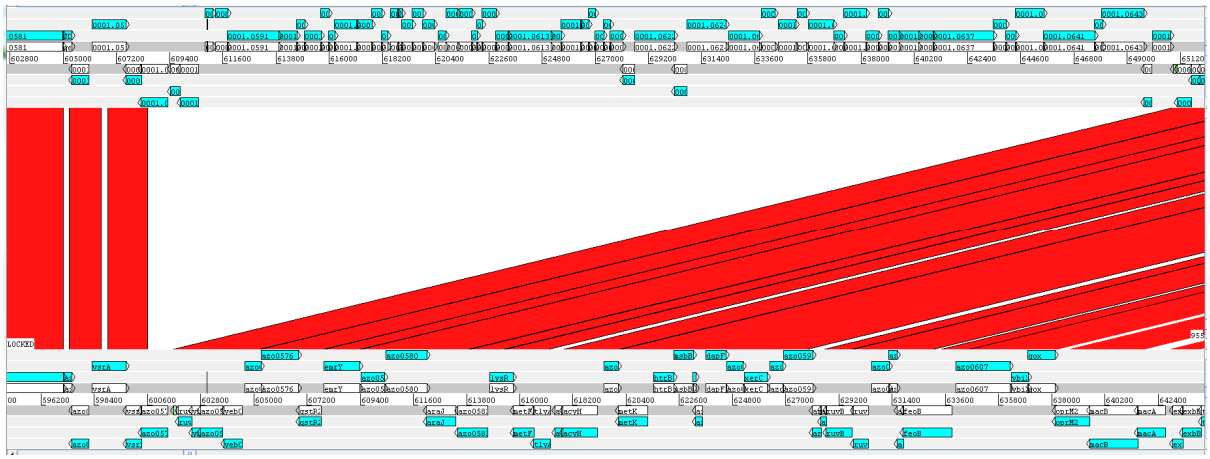

FIGURA 25 REGIÃO DE DIVERGÊNCIA 2
FONTE: O próprio autor através do programa ACT

TABELA 9 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 2

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.0587	transcriptional regulator [Streptomyces albus]		37
0001.0588	repressor [Paludibacterium yongneupense]		50
0001.0589	Nlp family transcriptional regulator [Thioalkalivibrio sulfidiphilus HL-EbGr7]		54
0001.0590	hypothetical protein Sputw3181_2932 [Shewanella sp. W3-18-1]	**Nenhum domínio conservado foi encontrado**	56
0001.0591	MuA-transposase/repressor protein CI, DNA-binding [Thauera sp. 63]		68
0001.0592	DNA transposition protein [Thauera sp. 63]		92
0001.0593	hypothetical protein [Vibrio nigripulchritudo]	**Nenhum domínio conservado foi encontrado**	45
0001.0594	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	60
0001.0595	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	66
0001.0597	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	38
0001.0598	hypothetical protein [Thauera sp. 28]	**Nenhum domínio conservado foi encontrado**	79
0001.0599	serine proteinase [Arthroderma otae CBS 113480]		38
0001.0601	single-stranded binding protein [Thauera sp. 63]	DNA sequence specific (IHF) and non-specific (HU) domains;	61
0001.0603	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	60
0001.0604	hypothetical protein [Thauera sp. 63]	Protein of unknown function (DUF1018);	79
0001.0605	hypothetical protein [Thauera sp. 63]	Mor transcription activator family; Mor (Middle operon regulator)	77
0001.0606	hypothetical protein [Streptococcus parauberis]	**Nenhum domínio conservado foi encontrado**	35
0001.0607	MULTISPECIES: hypothetical protein [Bacteroides]	F0F1 ATP synthase subunit B	52
0001.0608	N-acetylmuramoyl-L-alanine amidase, family 2 [Methyloversatilis universalis]	Peptidoglycan recognition proteins (PGRPs)	48
0001.0610	hypothetical protein [Leptothrix cholodnii]	**Nenhum domínio conservado foi encontrado**	57
0001.0611	hypothetical protein [Thauera phenylacetica]	**Nenhum domínio conservado foi encontrado**	49
0001.0612	conserved exported hypothetical protein [Candidatus Contendobacter odensis Run_B_J11]	**Nenhum domínio conservado foi encontrado**	45
0001.0613	hypothetical protein [uncultured bacterium A1Q1_fos_2101]	Mycoplasma protein of unknown function, DUF285	36
0001.0614	hypothetical protein [Pseudomonas mendocina]	**Nenhum domínio conservado foi encontrado**	40
0001.0615	RHS repeat-associated core domain protein [Leptospira interrogans serovar Bataviae str. UI 08561]	**Nenhum domínio conservado foi encontrado**	29
0001.0616	MULTISPECIES: hypothetical protein [Rhodocyclaceae]	**Nenhum domínio conservado foi encontrado**	37
0001.0617	hypothetical protein [Methyloversatilis universalis]	**Nenhum domínio conservado foi encontrado**	48
0001.0618	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	45
0001.0619	hypothetical protein [Methyloversatilis universalis]	**Nenhum domínio conservado foi encontrado**	58
0001.0620	hypothetical protein [Thauera sp. 63]	Protein of unknown function (DUF1804);	90
0001.0621	hypothetical protein [Halomonas sp. TD01]	**Nenhum domínio conservado foi encontrado**	30
0001.0622	phage protein [Thauera sp. 63]		89
0001.0623	hypothetical protein [Vibrio parahaemolyticus]	**Nenhum domínio conservado foi encontrado**	32

continua

TABELA 9 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 2

continuação

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.0624	hypothetical protein [Thauera sp. 63]	Protein of unknown function (DUF935); Mu-like prophage protein gp29	72
0001.0625	head morphogenesis protein spp1 gp7 [Thauera sp. 28]		70
0001.0626	phage virion morphogenesis protein, putative tail completion [Thauera sp. 63]		67
0001.0627	hypothetical protein [Geobacter sp. M18	ECF sigma factor;	68
0001.0628	putative membrane protein [Burkholderia sp. RPE67]	**Nenhum domínio conservado foi encontrado**	45
0001.0629	protease (I) and scaffold (Z) protein [Thauera sp. 63]		62
0001.0630	bacteriophage protein [Thauera sp. 28]	**Nenhum domínio conservado foi encontrado**	67
0001.0631	Mu-like prophage major head subunit gpT [Thauera sp. 28]		79
0001.0632	hypothetical protein [Thauera sp. 28]	**Nenhum domínio conservado foi encontrado**	51
0001.0633	hypothetical protein [Thauera sp. 63]	rotein of unknown function (DUF1320); Mu-like prophage protein gp36	67
0001.0634	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	54
0001.0635	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	69
0001.0636	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	52
0001.0637	phage-related minor tail protein [Thauera sp. 63]	tape measure domain; Not1 N-terminal domain; MAEBL;	44
0001.0638	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	55
0001.0639	hypothetical protein [Thioalkalivibrio sp. ALgr3]	**Nenhum domínio conservado foi encontrado**	43
0001.0640	hypothetical protein [Acinetobacter sp. CIP 101934]	WD40 domain; Domain of unknown function (DUF4613);	42
0001.0641	hypothetical protein [Thauera sp. 63]	DNA polymerase III subunits gamma and tau;	46
0001.0642	hypothetical protein [Thauera sp. 63]	**Nenhum domínio conservado foi encontrado**	37
0001.0643	hypothetical protein [Thioalkalivibrio sulfidiphilus]	**Nenhum domínio conservado foi encontrado**	36
0001.0644	hypothetical protein [Azoarcus sp. KH32C]	Helix-turn-helix XRE-family like proteins;	47
0001.0645	hypothetical protein [Thauera sp. 27]	Uncharacterized ACR, COG2135; DUF159 superfamily	52

FONTE: O próprio autor

5.6.3 Região 3

A terceira região adicional no genoma de A. *olearius* DQS4 (FIGURA 26) esta localizada entre as posições 1.092.696 e 1.134.832, possui 42.137 pb, codificando 28 proteínas (TABELA 10) com conteúdo GC de 59,5%.

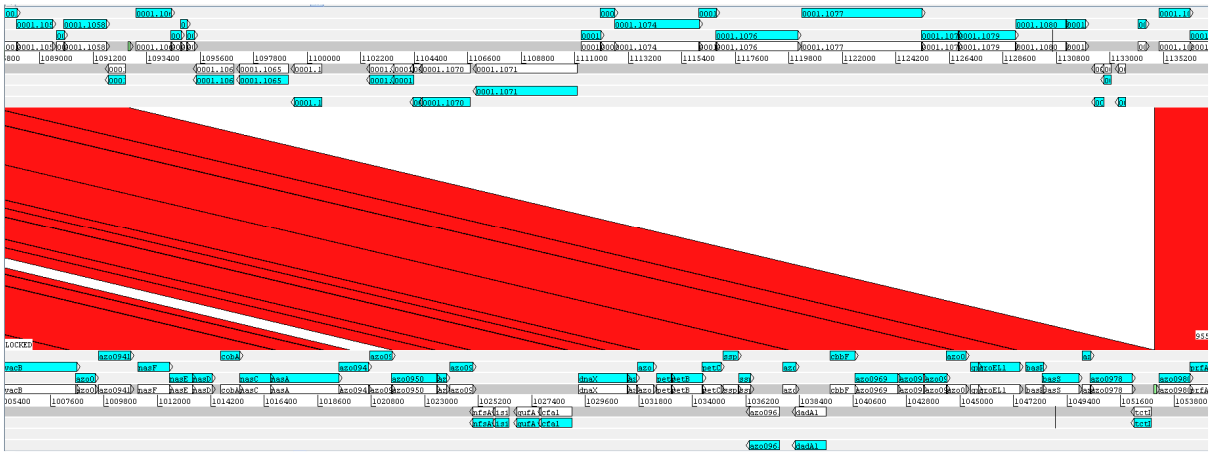

FIGURA 26 REGIÃO DE DIVERGÊNCIA 3
FONTE: O próprio autor através do programa ACT

TABELA 10 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 3

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.1060	Shufflon-specific DNA recombinase, putative [Ricinus communis]	Shufflon-specific DNA recombinase Rci and Bacteriophage Hp1_like integrase, C-terminal	50
0001.1061	hypothetical protein [Deefgea rivuli]	**Nenhum domínio conservado foi encontrado**	34
0001.1062	plasmid stabilization protein [Methylobacillus glycogenes]	Antitoxin of toxin-antitoxin stability system [Cell division and chromosome partitioning]; Antit	74
0001.1063	stability protein StbE [Thauera phenylacetica]	Cytotoxic translational repressor of toxin-antitoxin stability system [Translation, ribosomal	61
0001.1064	cell filamentation protein Fic [Oxalobacteraceae bacterium IMCC9480		71
0001.1065	putative Site-specific recombinase XerD [Thiomonas arsenitoxydans]	DNA breaking-rejoining enzymes, C-terminal catalytic domain; Predicted transcriptional reg	50
0001.1066	hypothetical protein [Burkholderia dilworthii]	**Nenhum domínio conservado foi encontrado**	59
0001.1067	hypothetical protein [Thauera sp. 28]	WYL domain; Predicted transcriptional regulator [Transcription]	97
0001.1068	ADP-ribosylglycohydrolase [Thauera sp. 28]		99
0001.1069	hypothetical protein [Thauera sp. 28]	**Nenhum domínio conservado foi encontrado**	100
0001.1070	hypothetical protein [Cupriavidus sp. amp6]	**Nenhum domínio conservado foi encontrado**	59
0001.1071	hypothetical protein [Cupriavidus sp. amp6]	**Nenhum domínio conservado foi encontrado**	62
0001.1072	hypothetical protein [Rubrivivax benzoatilyticus]	**Nenhum domínio conservado foi encontrado**	59
0001.1073	hypothetical protein [Burkholderia sp. A1]	Domain of unknown function (DUF1788	50
0001.1074	hypothetical protein [Rubrivivax benzoatilyticus	**Nenhum domínio conservado foi encontrado**	68
0001.1075	hypothetical protein SP5_022_00300 [Sphingomonas parapaucimobilis NBRC 15100]	**Nenhum domínio conservado foi encontrado**	33
0001.1076	restriction endonuclease subunit M [Rubrivivax benzoatilyticus]		67
0001.1077	hypothetical protein [Cupriavidus sp. UYPR2.512]	SPS1; AAA; AAA_12; STKc_PknB_like; NERD; COG3642;	68
0001.1078	hypothetical protein [Cupriavidus sp. UYPR2.512]	**Nenhum domínio conservado foi encontrado**	62
0001.1079	hypothetical protein [Lamprocystis purpurea]	TIGR02687 family protein	58
0001.1080	peptidase [Burkholderia sp. CCGE1001]	Putative ATP-dependent Lon protease; TIGR02688 family protein; Lon protease (S16) C-ter	88
0001.1081	hypothetical protein [Weeksella sp. FF8]	**Nenhum domínio conservado foi encontrado**	37
0001.1082	transposase [Burkholderia glumae]	HTH-like domain; HTH_21; PRK09409; IS2 transposase TnpB	64
0001.1083	Transposase [Desulfovibrio africanus	HTH_Tnp_1; HTH_29; HTH_Hin_like; Helix-turn-helix domain of Hin and related proteins; W	84
0001.1084	hypothetical protein [Burkholderia pseudomallei]	Prophage CP4-57 regulatory protein (AlpA); COG2452; Predicted site-specific integrase-re	63
0001.1085	unnamed protein product [uncultured bacterium]	Protein of unknown function (DUF3018);	56

FONTE: O próprio autor

5.6.4 Região 4

A região 4 (FIGURA 27) esta localizada entre as posições 1.895.144 e 1.909.492, possui 14.349 pb, codificando 24 proteínas (TABELA 11) com conteúdo GC é de 63,1%.

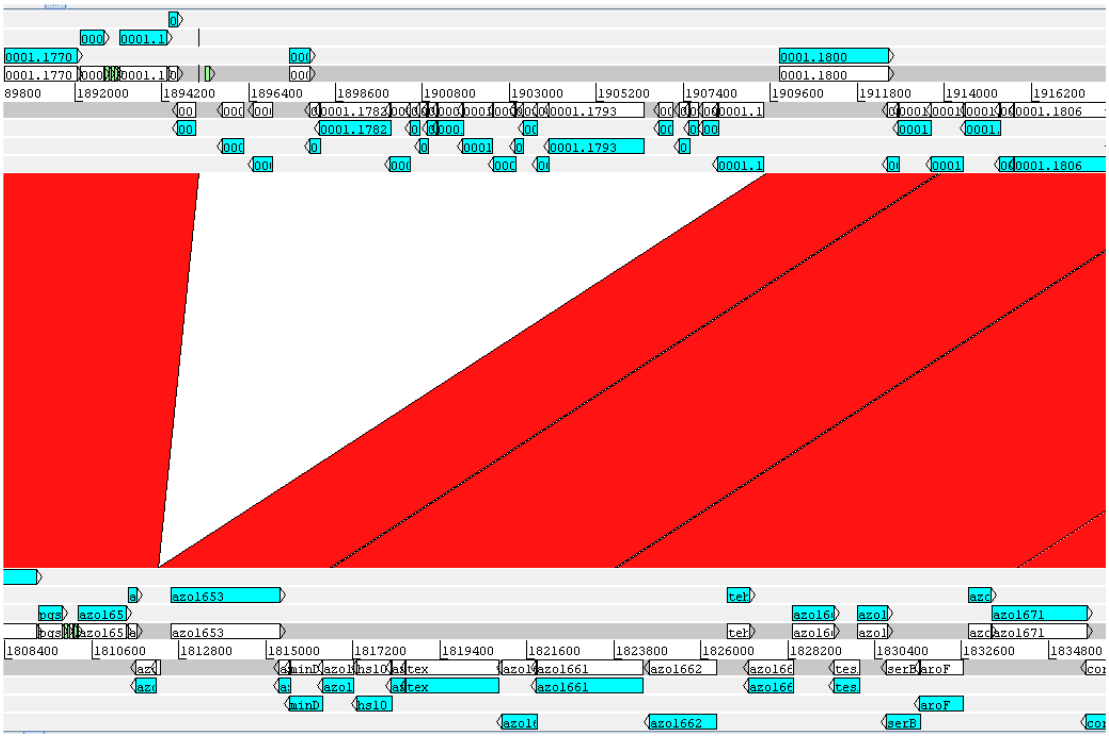

FIGURA 27 REGIÃO DE DE DIVERGÊNCIA 4
FONTE: O próprio autor através do programa ACT

TABELA 11 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 4

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.1778	hypothetical protein [Thauera phenylacetica]	DUF159; Uncharacterized ACR, COG2135; COG2135;	57
0001.1779	hypothetical protein [Burkholderia sp. Ch1-1]	**Nenhum domínio conservado foi encontrado**	45
0001.1780	hypothetical protein [SAR324 cluster bacterium SCGC AAA240-J09]	Short C-terminal domain;	42
0001.1781	biotin-requiring enzyme [Leptospira interrogans]		35
0001.1782	terminase, partial [Kiloniella laminariae]		53
0001.1783	hypothetical protein [Burkholderia cepacia]	**Nenhum domínio conservado foi encontrado**	33
0001.1784	FUN14 family protein [Natronorubrum sulfidifaciens]		40
0001.1785	hypothetical protein H257_15265 [Aphanomyces astaci]	**Nenhum domínio conservado foi encontrado**	52
0001.1786	hypothetical protein Ao3042_03040 [Aspergillus oryzae 3.042]	**Nenhum domínio conservado foi encontrado**	46
0001.1787	hypothetical protein [Pseudomonas aeruginosa]	**Nenhum domínio conservado foi encontrado**	29
0001.1788	segregation protein A [Desulfonatronovibrio hydrogenovorans]		43
0001.1789	diguanylate cyclase [Exiguobacterium sp. NG55]		31
0001.1790	PREDICTED: B3 domain-containing protein At5g42700-like isoform X1 [Solanum tuberosum]		39
0001.1791	excisionase [Leucobacter sp. UCD-THU]	GT1 family of glycosyltransferases. ExpC; Helix-turn-helix domain	56
0001.1792	XRE family transcriptional regulator [Xanthobacter autotrophicus]		32
0001.1793	RecA-family ATPase (modular protein) [Mesorhizobium metallidurans]	AAA domain; Hexameric Replicative Helicase RepA; RecA-family ATPase	30
0001.1794	hypothetical protein [Laribacter hongkongensis]	PHA00675;	48
0001.1796	hypothetical protein [Thiomonas arsenitoxydans]	**Nenhum domínio conservado foi encontrado**	47
0001.1797	hypothetical protein [Cupriavidus basilensis]	**Nenhum domínio conservado foi encontrado**	61
0001.1798	hypothetical protein MYCTH_2298982 [Myceliophthora thermophila ATCC 42464]	**Nenhum domínio conservado foi encontrado**	39
0001.1799	hypothetical protein [Cupriavidus sp. UYPR2.512]	XerC Integrase [DNA replication, recombination, and repair];	66

FONTE: O próprio autor

5.6.5 Região 7

A região 7 (FIGURA 28) esta localizada entre as posições 2.687.896 e 2.708.263, possui 20.368 pb, codificando 19 proteínas (TABELA 12) e com conteúdo GC é de 69,3%. Próximo a esta região também foram encontrados dois genes que codificam uma proteína sulfurtransferase, a primeira com 82% de identidade com o organismo *Thauera linaloolentis* e a segunda com 80% de identidade com o organismo *Janthinobacterium* sp. *Marseille*, a terceira é uma proteína de membrana com 64% de identidade com o mesmo organismo anterior. Ambas as proteínas não existem no genoma do *Azoarcus* BH72.

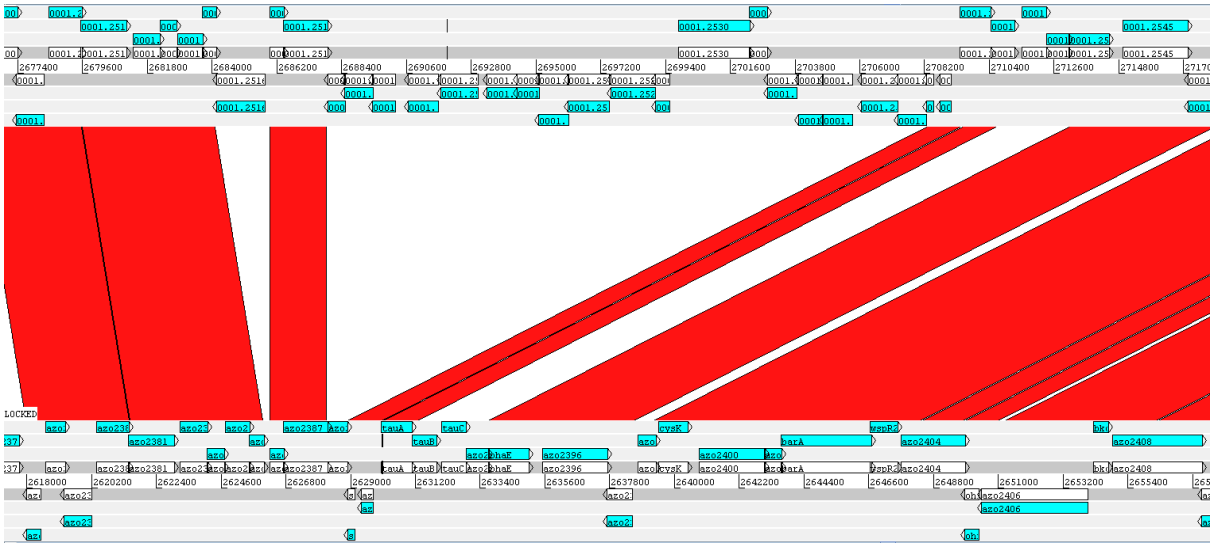

FIGURA 28 REGIÃO DE DIVERGÊNCIA 7
FONTE: O próprio autor através do programa ACT

TABELA 12 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 7

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.2519	cysteine dioxygenase type I [Thauera linaloolentis]		87
0001.2520	2-dehydropantoate 2-reductase [Azotobacter vinelandii]		68
0001.2521	aldolase [Bordetella trematum]		51
0001.2522	MFS transporter [Rhodopseudomonas palustris]		67
0001.2523	alkanesulfonate monooxygenase [Burkholderia sp. WSM4176]		55
0001.2524	ABC transporter substrate-binding protein [Pseudomonas thermotolerans]	Periplasmic_Binding_Protein_Type_2; Aromatic solutes transporter of Bug	74
0001.2525	Enoyl-CoA hydratase/isomerase [Rhodopseudomonas palustris]		70
0001.2526	glycine reductase [Bradyrhizobium sp. YR681]		67
0001.2527	phenylacetate--CoA ligase [Rhodopseudomonas palustris]		72
0001.2528	hypothetical protein [Azotobacter vinelandii]	TctA; COG3333; Tripartite tricarboxylate transporter TctA family;	86
0001.2529	hypothetical protein [Azotobacter vinelandii]	TctB; Tripartite tricarboxylate transporter TctB family	64
0001.2530	biotin transporter BioY [Ottowia thiooxydans]		59
0001.2531	hypothetical protein [Azoarcus sp. BH72]	**Nenhum domínio conservado foi encontrado**	94
0001.2532	sulfonate ABC transporter ATP-binding protein [Curvibacter lanceolatus]		69
0001.2533	ABC-type transporter membrane permease [Thauera sp. 27]		87
0001.2534	ABC transporter substrate-binding protein [Variovorax paradoxus]		67
0001.2535	alkanesulfonate monooxygenase [Thauera linaloolentis]		82
0001.2536	aliphatic sulfonate ABC transporter substrate-binding protein [Thauera sp. 27]		82

FONTE: O próprio autor

5.6.6 Região 8

A região 8 (FIGURA 29) esta localizada entre as posições 2.817.041 e 2.826.863, possui 9.823 pb, codificando 10 proteínas (TABELA 13) e com conteúdo GC de 68,8%.

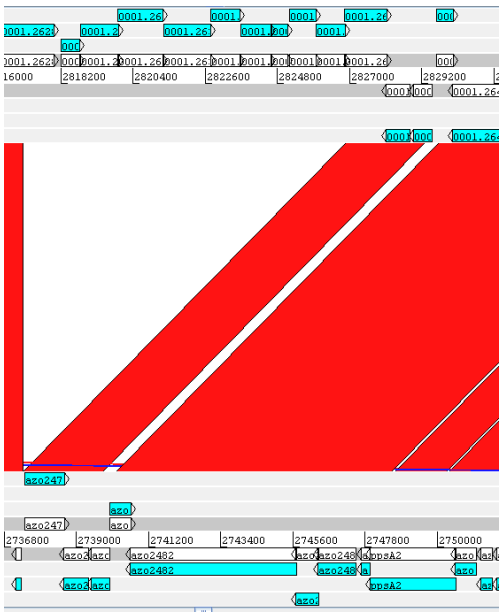

FIGURA 29 REGIÃO DE DE DIVERGÊNCIA 8
FONTE: O próprio autor através do programa ACT

TABELA 13 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 8

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.02629	peroxiredoxin [Pandoraea sp. B-6]		84
0001.02630	acyl-CoA dehydrogenase [Burkholderia cepacia]		75
0001.02631	rubredoxin-type Fe(Cys)4 protein [Herbaspirillum frisingense]	Pyr_redox_2	64
0001.02632	ABC transporter nitrate-binding protein [Pseudomonas sp. BAY1663]		77
0001.02633	nitrate ABC transporter permease [Herbaspirillum frisingense]		81
0001.02634	nitrate ABC transporter ATPase [Herbaspirillum frisingense]		82
0001.02635	cyanate hydratase [Pseudomonas sp. BAY1663]		89
0001.02636	hypothetical protein [Pseudomonas sp. CF149]	Sulfite exporter TauE/SafE	57
0001.02637	beta-lactamase [Ralstonia sp. PBA]		70

FONTE: O próprio autor

5.6.7 Região 10

A região 10 (FIGURA 30) esta localizada entre as posições 3.766.629 e 3.778.818, possui 12.190 pb, codificando 8 proteínas (TABELA 14). Seu conteúdo GC é de 62,2%.

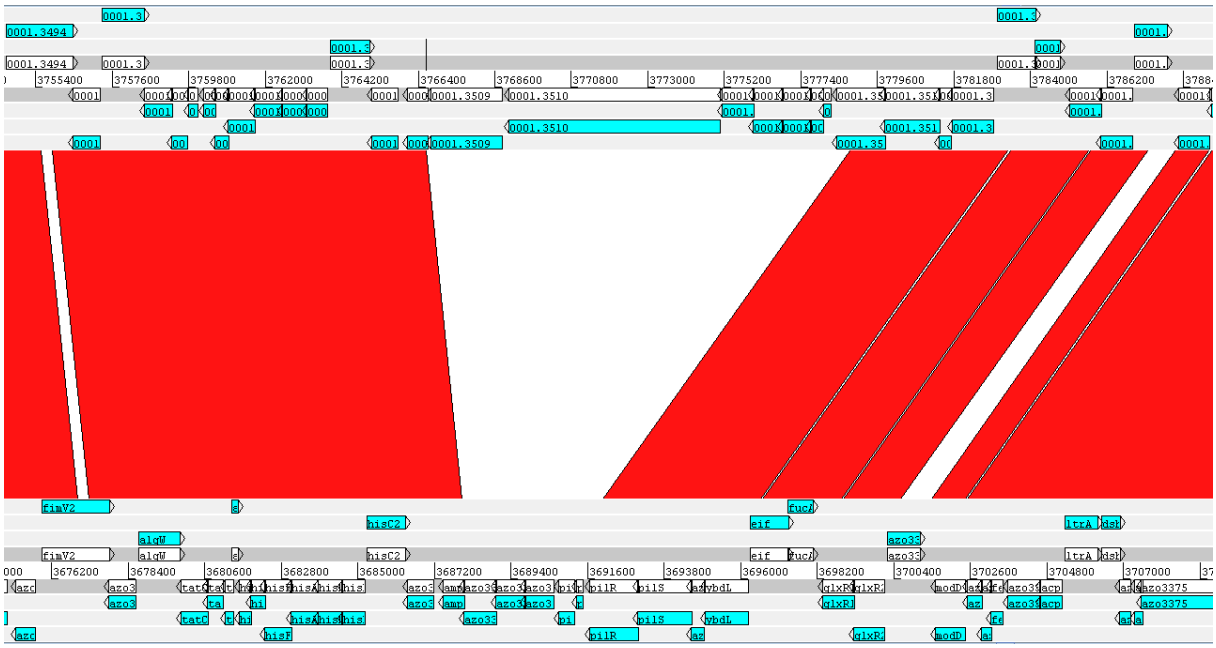

FIGURA 30 REGIÃO DE DE DIVERGÊNCIA 10
FONTE: O próprio autor através do programa ACT

TABELA 14 GENES CODIFICANTES NA REGIÃO DE DIVERGÊNCIA 10

CDS	Proteína / Nome do organismo	Domínio Conservado de Proteínas Hipotéticas	Identidade %
0001.3509	sulfatase [Rubrobacter xylanophilus]		39
0001.3510	hypothetical protein [Pelodictyon phaeoclathratiforme]	cellulose synthase subunit BcsC; Tetratricopeptide repeat domain;	35
0001.3511	putative ABC transporter ATP-binding protein YxlF [Candidatus Accumulibacter sp. SK-02]		60
0001.3512	hypothetical protein [Deefgea rivuli]	**Nenhum domínio conservado foi encontrado**	44
0001.3513	ABC-type transport system [Candidatus Accumulibacter sp. SK-12]		52
0001.3514	penicillin-binding protein [Candidatus Accumulibacter phosphatis]		41
0001.3515	pilin domain-containing protein [Pseudomonas stutzeri]		77

FONTE: O próprio autor

5.7 GENES DE INTERESSE ENCONTRADOS

O principal interesse na bactéria foi a busca pelos genes de metabolismo de nitrogênio e metabolismo de componentes aromáticos. No processo de análise do genoma montado foram encontrados outros genes que chamaram atenção, os quais podem ser destacados o metabolismo de enxofre, produção de sideróforos e gene da proteína transportadora de fosfato que perdeu sua função.

5.7.1 Genes de metabolismo de nitrogênio

Todos os genes relacionandos com o metabolismo de nitrogênio do *Azoarcus* BH72 foram encontrados no *A. olearius* DQS4 como pode ser visualizado na (TABELA 15).

TABELA 15 COMPARAÇÃO DOS GENES DO CLUSTER DE FIXAÇÃO DE NITROGÊNIO

Organismos	A. DQS4	A. BH72
NifA	X	X
NrpR		
CysE		
NifS	X	X
NifU	X	X
IscA-like	X	X
NifB	X	X
frdN	X	X
NifX	X	X
NifX2	X	X
NafY	X	X
NifB-1		
NifB-2		
NifE	X	X
NifN	X	X
NifQ	X	X
NifV	X	X
NifV-a		
NifV-o		
NifW	X	X
NifM	X	X
NifH	X	X
NifD	X	X
NifK	X	X
NifZ	X	X
NifT		
NifO	X	X
Avin2460	X	X
NifY	X	X

FONTE: O próprio autor

Foi realizada a comparação de todos os genes relacionados com fixação de nitrogênio do *A. olearius* DQS4 com os demais genomas de *Azoarcus* disponíveis no banco de dados do NCBI: *Azoarcus* BH72, *Azoarcus* KH32C, *A. toluclasticus* ATCC700605 e *Azoarcus* EbN1 (TABELA 16). A comparação entre o cluster Nif do *A. olearius* DQS4 e o *Azoarcus* BH72 mostrou que os genes que fazem parte desse sistema são muito semelhantes, sendo que a maioria dos genes é idêntica. Nos demais genomas o que mais se identifica com o DQS4 é o A. *toluclasticus* ATCC700605 e o menos semelhante foi o *Azoarcus* EbN1 que praticamente não possuem genes relacionados com fixação de nitrogênio por ser uma bactéria desnitrificadora, transformando nitritos e nitratos em nitrogênio (RABUS et al, 2005).

TABELA 16 COMPARAÇÃO DOS GENES DE FIXAÇÃO DE NITROGÊNIO ENTRE ORGANISMOS DO GÊNERO AZOARCUS

DQS-4T		BH72			KH32C			ATCC700605			EbN1		
Nome do gene	Função	Hit	Posição	% identidade	Hit	Posição	% identidade	Hit	Posição	% identidade	Hit	Posição	% identidade
NifL	nitrogen fixation negative regulator	bi	524	100	bi	3364	37.07	bi	3895	40.8	uni	3874	31.62
NifA	Nitrogenase-specific transcriptional regulator	bi	525	100	uni	1043	50.82	uni	194	44.8	uni	3297	45.56
0001.0527	hypothetical protein	bi	526	99.23	-	-	-	-	-	-	-	-	-
0001.0528	hypothetical secreted protein	bi	527	98.18	-	-	-	-	-	-	-	-	-
SodC	Superoxide dismutase [Cu-Zn] precursor	bi	528	100	bi	4405	37.41	bi	4697	37.87	-	-	-
0001.0530	hypothetical protein	bi	529	100	-	-	-	-	-	-	-	-	-
NifB	Nitrogenase FeMo-cofactor synthesis FeS core	bi	530	100	bi	4006	58.89	bi	4174	53.39	uni	3120	23.74
0001.0532	4Fe-4S ferredoxin, iron-sulfur binding	bi	531	100	uni	2857	55	bi	96	57.41	uni	3090	53.33
NifO	Nitrogenase-associated protein NifO	bi	532	100	-	-	-	-	-	-	-	-	-
0001.0534	beta-lactamase domain protein	bi	533	99.77	uni	3890	28.29	uni	4999	29.61	uni	280	34.3
FdxC	Ferredoxin	bi	534	100	-	-	-	uni	1742	28.26	-	-	-
FdxD	Ferredoxin	bi	535	100	bi	791	41.88	-	-	-	-	-	-
NifQ	Nitrogenase FeMo-cofactor synthesis molybdenum delivery	bi	536	100	bi	4033	48.98	bi	4204	43.33	-	-	-
DraG1	ADP-ribosylglycohydrolase	bi	537	99.69	bi	797	45.17	bi	4180	48.62	-	-	-
0001.0539	hypothetical protein	bi	538	99.03	-	-	-	bi	2145	73.08	bi	603	52.25
0001.0540	hypothetical protein	bi	539	100	-	-	-	bi	2146	83.9	bi	602	30.77
0001.0541	Inner membrane protein yjdF	bi	540	97.54	-	-	-	-	-	-	-	-	-
0001.0542	hypothetical protein	bi	541	98.64	bi	3889	36.55	bi	5000	39	bi	2218	33.86
0001.0543	hemerythrin-like metal-binding protein	bi	542	100	uni	1232	27.27	uni	2421	38.93	-	-	-
0001.0544	hypothetical protein	bi	543	100	bi	798	31.39	-	-	-	-	-	-
DraT	NAD(+)--dinitrogen-reductase ADP-D-ribosyltransferase	bi	544	99.64	bi	796	41.73	bi	4181	42.42	-	-	-
NifH	Nitrogenase (molybdenum-iron) reductase and maturation	bi	545	100	bi	794	78.17	bi	4183	76.31	uni	2521	41.67
NifD	Nitrogenase (molybdenum-iron) alpha chain	bi	546	100	bi	4025	74.74	bi	4184	72.92	-	-	-
NifK	Nitrogenase (molybdenum-iron) beta chain	bi	547	100	bi	792	56.32	bi	4185	56.56	-	-	-
NifT	NifT/FixU	bi	548	100	bi	799	47.06	bi	4177	42.65	-	-	-
FdxN	4Fe-4S ferredoxin, iron-sulfur binding	bi	549	100	uni	4007	53.97	bi	4207	75.38	uni	3090	49.25
NifY1	NifY protein	bi	550	99.58	uni	784	34.56	uni	4200	37.1	-	-	-
0001.0552	hypothetical protein	bi	551	100	bi	4015	34.94	-	-	-	-	-	-
0001.0553	chemotaxis sensory transducer	bi	552	100	uni	2980	41.73	uni	925	44.13	uni	3583	32.48
NifM	NifM protein	bi	553	100	uni	1779	34.38	uni	3129	32.45	uni	3839	34.59
NifZ	NifZ protein	bi	554	100	bi	4010	48.65	bi	4176	45.95	-	-	-
NifW	Nitrogenase stabilizing/protective protein NifW	bi	555	100	bi	4049	49.5	bi	4216	45.1	-	-	-
0001.0557	hypothetical protein	bi	556	100	-	-	-	-	-	-	-	-	-
NifP	Serine acetyltransferase	bi	557	99.65	bi	4050	65.55	uni	773	51.75	uni	3696	51.32
NifV	Homocitrate synthase	bi	558	100	bi	776	58.89	bi	4217	57.18	uni	4102	36
NifS	Cysteine desulfurase, NifS subfamily	bi	559	100	bi	4013	46.13	uni	770	44.56	uni	3693	43.52
NifU	Iron-sulfur cluster assembly scaffold protein NifU	bi	560	100	uni	2772	50.82	uni	769	50.82	uni	3692	48.8
HesB	iron binding protein	bi	561	100	uni	1059	46.73	uni	3923	44.86	uni	505	48.96
0001.0563	hypothetical protein	bi	562	100	-	-	-	-	-	-	-	-	-
FdxB	4Fe-4S ferredoxin, nitrogenase-associated	bi	563	100	bi	4032	39	bi	4203	35.79	-	-	-
0001.0565	hypothetical protein	bi	564	100	bi	782	44.07	bi	4202	42.42	-	-	-
0001.0566	NifX-associated protein	bi	565	100	bi	4030	48.2	bi	4201	34.72	-	-	-
0001.0567	hypothetical sigma-E factor regulatory protein	bi	566	98.52	-	-	-	-	-	-	-	-	-
NifX	Nitrogenase FeMo-cofactor carrier protein NifX	bi	567	100	bi	784	39.5	bi	4200	38.66	-	-	-
NifN	Nitrogenase FeMo-cofactor scaffold and assembly	bi	568	99.57	bi	785	46.31	bi	4199	44.89	-	-	-
NifE	Nitrogenase FeMo-cofactor scaffold and assembly	bi	569	100	bi	4027	55.16	bi	4197	51.4	-	-	-

FONTE: O próprio autor utilizando informações do programa RAST

O cluster Nif do *A. olearius* DQS4 fica localizado entre as posições 560.887 e 599.405 do seu genoma, iniciando com o gene *nifL* até o *nifE*. Na comparação das vias de metabolismo de nitrogênio entre os genomas de DQS4 e BH72 foi confirmado que o DQS4 possui um gene codificando uma cianato hidratase (FIGURA 31), o qual o BH72 não possui (FIGURA 32). Essa enzima é responsável por realizar a conversão de cianeto em amômia que também pode ser utilizada em processos metabólicos da bactéria. Este gene esta localizado na região de inserção de divergência 8.

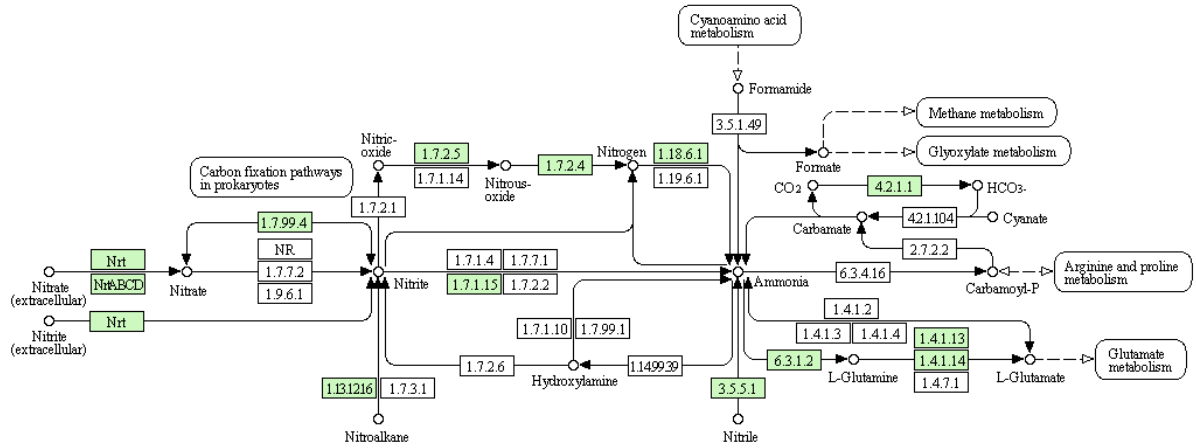

FIGURA 31 METABOLISMO DE NITROGENIO NO *Azoarcus* BH72

FONTE: O próprio autor através do programa KEEG

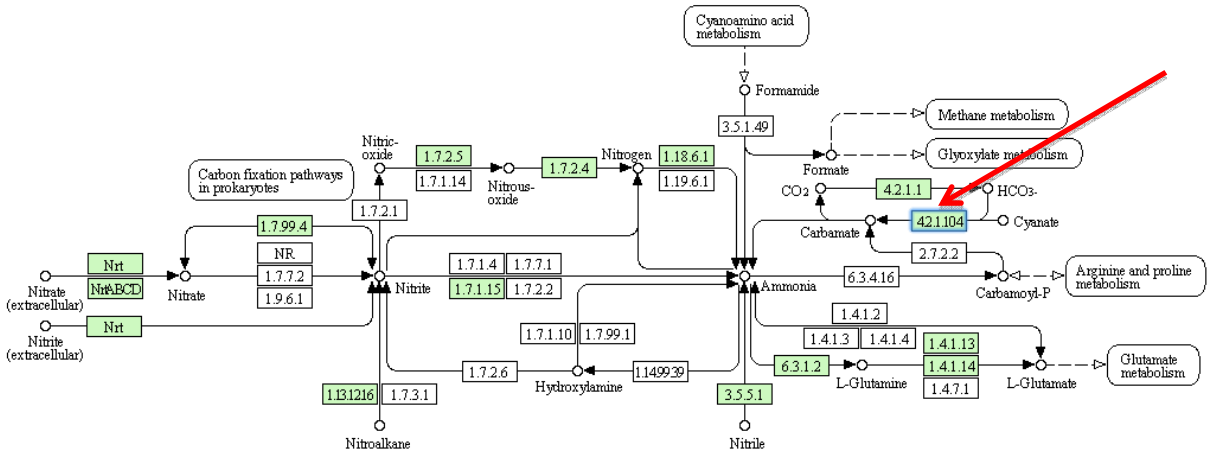

FIGURA 32 METABOLISMO DE NITROGENIO DO *A. olearius* DQS4

FONTE: O próprio autor através do programa KEEG

5.7.2 Genes de metabolismo de componentes aromáticos

Os genes responsáveis pelo metabolismo de componentes aromáticos são exatamente os mesmo genes encontrados no *Azoarcus* BH72, não havendo inserção de novos genes.

5.7.3 Genes de associação com plantas e promoção do crescimento vegetal

A enzima 1-aminocyclopropane-1-carboxylate (ACC) deaminase promotora do crescimento vegetal, responsável por baixar o nível de etileno na planta, não foi encontada no genoma do *A. olearius* DQS4, assim como não foi encontrado nenhum gene relacionado ao sistema de secreção do tipo 3, geralmente relacionado à associação com plantas. Também foram encontrados alguns genes associados com a produção de sideróforos após a região de inserção de numero 8, os quais o BH72 não possui. O BH72 também não possui a enzima (ACC) deaminase.

5.7.4 Genes de metabolismo de enxofre

Foram encontrados três genes relacionados com o metabolismo de enxofre na região de divergência 7 do DQS4 e que o BH72 não apresenta (FIGURAS 33 e 34). Especificamente, essas proteínas estão relacionadadas com assimilação e utilização de enxofre orgânico na forma de alkanosulfatos. De maneira semelhante, o genoma de *Pseudomoas aeruginosa* N002, isolada de solo contaminado com óleo, apresentou alta similaridade com o genoma de *P. aeruginosa* JN661695. Entretando, a estirpe N002 contém mais genes relacionados com degradação de hidrocarbonetos, incluindo alkanosulfonato monoxigenase (ROY et al, 2013).

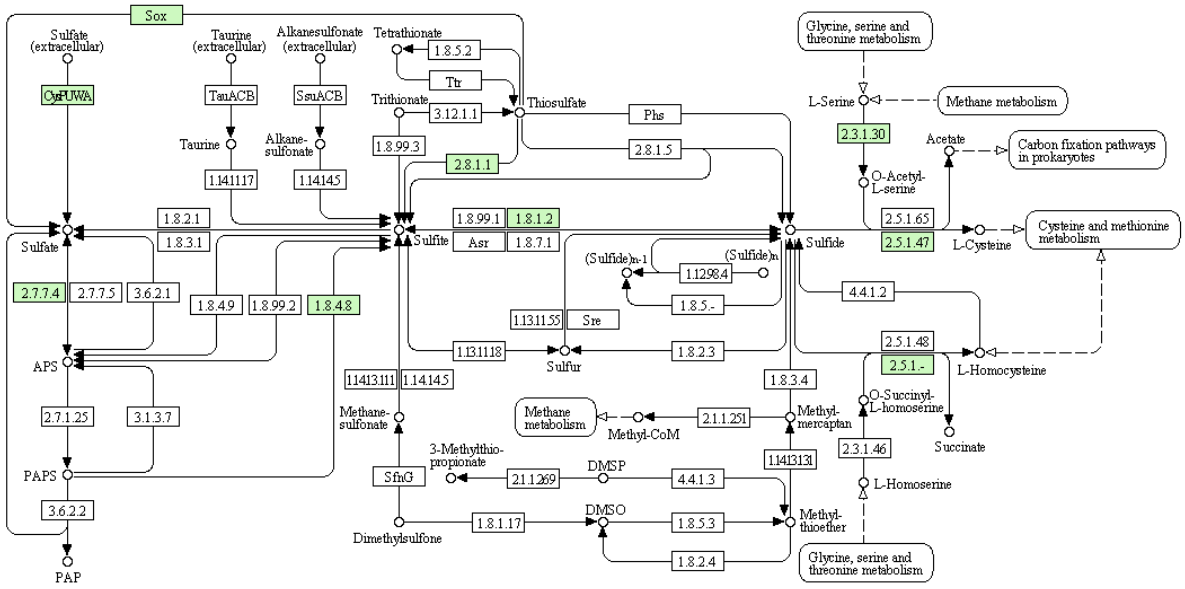

FIGURA 33 METABOLISMO DE ENXOFRE DO AZOARCUS BH72

FONTE: O próprio autor através do programa KEEG

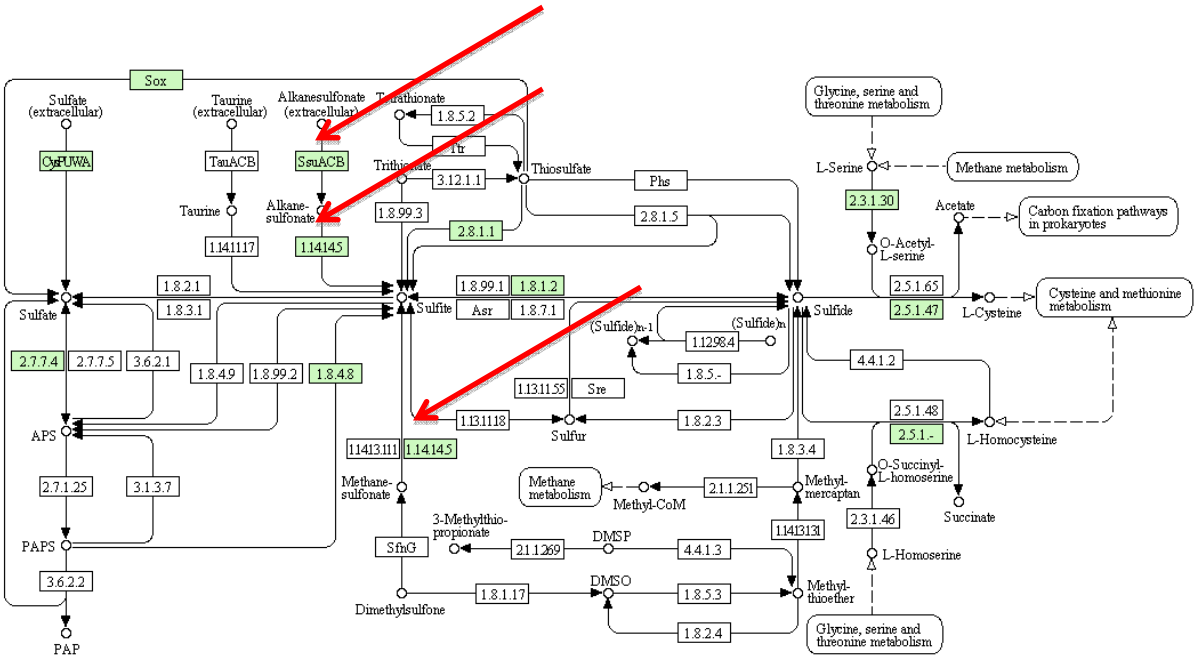

FIGURA 34 METABOLISMO DE ENXOFRE DO AZOARCUS DQS4

FONTE: O próprio autor através do programa KEEG

5.7.5 Proteína transportadora de fosfato

No processo de anotação do genoma foi detectada uma inserção de uma base "G" Guanina, no interior do gene da proteína transportadora de fosfato (*phosphate transporter* PitA, NCBI *Reference Sequence*: WP_011766380.1) o qual produziu um *frameshift*. Localizado na posição 3.022.662 do genoma até a base 3.024.143, conforme imagem retirada do programa Artemis. Como pode ser localizado na (FIGURA 35), onde existe uma queda no conteúdo GC da proteína, simbolizado pela linha azul no gráfico.

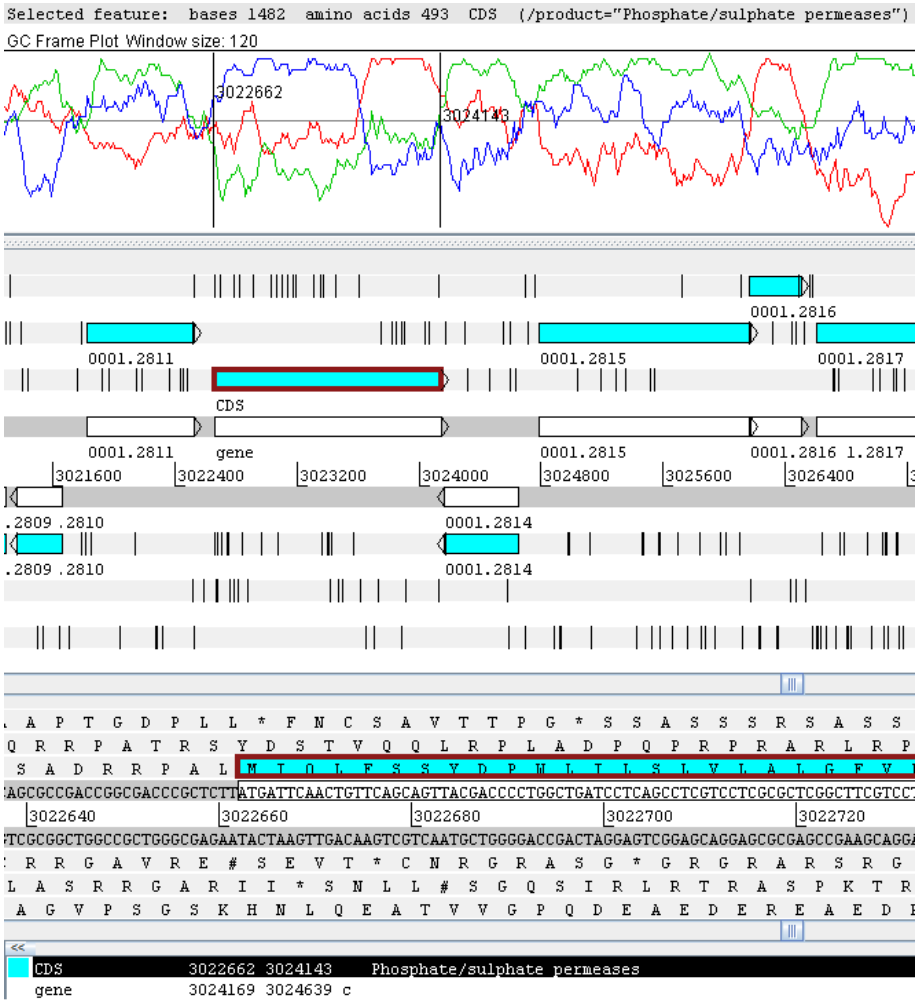

FIGURA 35 PROTEÍNA DE FOSFATO

FONTE: O próprio autor através do programa Artemis.

Conforme as imagens geradas pelo programa interpro, disponível em <http://www.ebi.ac.uk/interpro/> é possível inferir com estes dados que a proteína codificada perdeu ou modificou sua função, pois a região C-terminal perdeu os domínios transmembrana e passou a ter um possível domínio citoplasmático (FIGURAS 36 e 37).

FIGURA 36 PROTEÍNA TRANSPORTADORA DE FOSFADO DE *Azoarcus* BH72
FONTE: O próprio autor através do programa Interpro

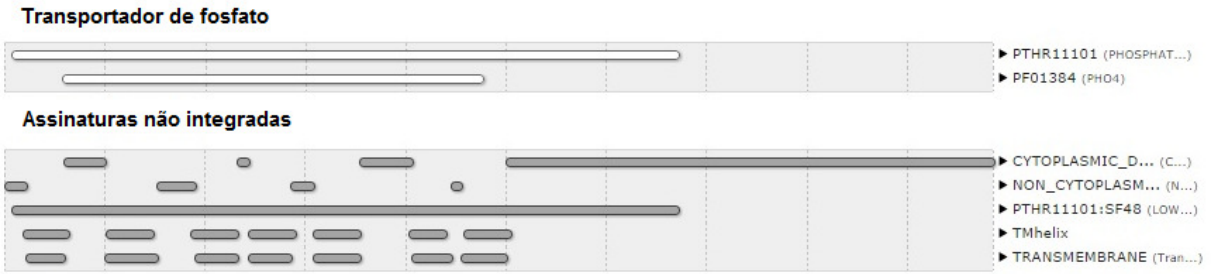

FIGURA 37 PROTEÍNA TRANSPORTADORA DE FOSFATO DE DE *A. olearius* DQS4
FONTE: O próprio autor através do programa Interpro

Segundo GEBHARD e colaboradores (2009) esta proteína é dispensavel para o crescimento da bactéria *Mycobacterium smegmatis*, porém esta proteína é importante na *Escherichia coli* para manter os níveis de fosfato inorgânico e zinco (ACKSON et al, 2009).

6. CONCLUSÕES

- O genoma de A. *olearius* DQS4 foi completamente sequenciado e fechado;
- Análises comparativas com outros genomas de bactérias do gênero Azoarcus sugerem que A. *olearius* DQS4 e *Azoarcus* BH72 podem ser classificados como estirpes diferentes da mesma espécie;
- *Olearius* DQS4 apresenta um sistema para transporte e metabolismo de cianato e alkanosufonatos que está ausente na estirpe BH72;
- Mesmo sendo organismos isolados de locais e ambientes diferentes, A. *olearius* DQS4 de um ambiente contaminado com óleo (Chen et al., 2013) e *Azoarcus* BH72 de raízes da gramínea *Kallar* (Reinhold-Hurek et al., 1993a), ambos possuem uma alta similariedade em seus genomas e compartilham de muitas genes em comum, principalmente genes envolvidos com a colonização e promoção de crescimento vegetal.

REFERÊNCIAS

ALTSCHUL, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. Nucleic Acids Res 25: 3389–402. 1997.

ALVES, M., e GEHLEN, C. "**Mapeamento de genes nif publicados no ncbi usando conceitos de mineração de dados e inteligência artificial mapeamento de genes nif publicados no ncbi usando conceitos de mineração de dados e inteligência**. 2011.

ANDERS, H. J., KAETZKE, A., KÄMPFER, P., LUDWIG, W. e FUCHS, G. **Taxonomic position of aromatic-degrading denitrifying pseudomonad strains K 172 and KB 740 and their description as new members of the genera Thauera, as Thauera aromatica sp. nov., and Azoarcus, as Azoarcus evansii sp. nov., respectively, members of the beta subclass of the Proteobacteria**. Int. J. Syst. Bacteriol 45, 327-333. 1995.

AZIZ, R. K., BARTELS, D., BEST, A. A., DEJONGH, M., DISZ, T., EDWARDS, R. A., FORMSMA, K., GERDES, S., GLASS, E. M., KUBAL, M., MEYER, F., OLSEN, G. J., OLSON, R., OSTERMAN, A. L., OVERBEEK, R. A., MCNEIL, L. K., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G. D., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A., ZAGNITKO, O. **The RAST Server: Rapid Annotations using Subsystems Technology.** BMC Genomics, 2008.

BALDANI, V. L., BALDANI D., DOBEREINER, J. **Host-plant specificity in the infection of cereals with Azospirillum spp. Programa Fixação Biologica de Nitrogênio**, SNLCS/EMBRAPA, CNPq, Km 47, 23460 Seropédica, Rio de Janeiro, Brazil. 1979.

CARVER T, BERRIMAN M, TIVEY A, PATEL C, BOHME U, **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database**. Bioinformatics 24: 2672–2676. 2008.

CHAISSON, M. J., e PEVZNER, P. A. **Short read fragment assembly of bacterial genomes.** Genome Research, 18(2), 324–330. 2008.

CHEN, M.-H., SHEU, S.-Y., JAMES, E. K., YOUNG, C.-C., e CHEN, W.-M. **Azoarcus olearius sp. nov., a nitrogen-fixing bacterium isolated from oil-contaminated soil**. International journal of systematic and evolutionary microbiology, ijs.0.050609–0–. doi:10.1099/ijs.0.050609-0. 2013.

DOBEREINER, J. **A importância da fixação biológica de nitrogênio para a agricultura sustentável. Biotecnologia Ciência**, 48–49. Available at: http://www.agencia.cnptia.embrapa.br/Repositorio/revistabiotecnologia1ID-0wLrpPCbk3.pdf (accessed 16/02/14). 1997.

DOS SANTOS, P. C., FANG, Z., MASON, S. W., SETUBAL, J. C. e DIXON, R. **"Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes.,"** BMC Genomics, vol. 13, p. 162, Jan. 2012.

EWING, B., HILLIER, L., WENDL, M. C., GREEN, P. **Base-Calling of Automated Sequencer Traces Using Phred . I . Accuracy Assessment. Genome research**, p. 175–185, 1998.

GENBANK. Disponível em **<http://www.ncbi.nlm.nih.gov/genbank/>** Acessado em 23/09/2013.

HUREK, T., REINHOLD-HUREK, B. **Azoarcus sp. strain BH72 as a model for nitrogen-fixing grass endophytes**. Journal of Biotechnology, 106(2-3), 169–178. 2003.

KANEHISA, M., GOTO, S., HATTORI, M., AOKI-KINOSHITA, K. F., ITOH, M., **From genomics to chemical genomics: new developments in KEGG**. Nucleic Acids Res 34: D354–357. 2006.

KNUDSEN, T. B., M. FLENSBORG, M. **"CLC Genomics Workbench, 4.0,"** 2008.
KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C., SALZBERG, S. L., **Versatile and open software for comparing large genomes**. Genome Biol., v. 5:R12, 2004.

LAGESEN K,. **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. Nucleic Acids Res. 35:3100–3108. 2007.

LOWE, T. M., EDDY, S. R. **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. Nucleic Acids Res 25: 955–964. 1997.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J. HE, G., CHEN, Y., PAN, Q., LIU, Y. TANG, J., WU, G., ZHANG, H., SHI, Y., LIU, Y., YU, C., WANG, B., LU, Y., HAN, C., CHEUNG, W. D., YIU, S.-M., PENG, S., XIAOQIAN, Z., LIU, G., LIAO, X., LI, Y., YANG, H., WANG, J., LAM, T.-W., e WANG, J. **"SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.,"** Gigascience, vol. 1, p. 18, 2012.

MECHICHI, T., STACKEBRANDT, E., GAD'ON, N. e FUCHS, G. **Phylogenetic and metabolic diversity of bacteria degrading aromatic compounds under denitrifying conditions, and description of Thauera phenylacetica sp. Nov., Thauera aminoaromatica sp. nov., and Azoarcus buckelii sp. nov**. Arch. Microbiol 178, 26-35. 2002.

MILLER,J.R. **Aggressive assembly of pyrosequencing reads with mates. Bioinformatics**, 24, 2818–2824. 2008.

MORIYA, Y., ITOH, M., OKUDA, S., YOSHIZAWA, A., AND KANEHISA, M.; **KAAS: an automatic genome annotation and pathway reconstruction server**. Nucleic Acids Res. 35, W182-W185. 2007.

MUTHUKUMARASAMY, R., REVATHI, G., SESHADRI, S., e LAKSHMINARASIMHAN, C. (2002). **Gluconacetobacter diazotrophicus (syn.Acetobacter diazotrophicus), a promising diazotrophic endophyte in tropics**. Current Science, 83(2), 137–145.

NISHIZAWA, T., TAGO, K., OSHIMA, K., HATTORI, M., ISHII, S., OTSUKA, S., e SENOO, K. **Complete genome sequence of the denitrifying and N2O-reducing bacterium Azoarcus sp. strain KH32C**. Journal of bacteriology, 194(5), 1255. doi:10.1128/JB.06618-11. 2012.

OVERBEEK R, OLSON R, PUSCH GD, OLSEN GJ, DAVIS JJ, DISZ T, EDWARDS RA, GERDES S, PARRELLO B, SHUKLA M, VONSTEIN V, WATTAM AR, XIA F, STEVENS R. **The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).** Nucleic Acids Res. 2014.

PARADA, M. **Sinorhizobium fredii HH103 mutants affected in capsular polysaccharide (KPS) are impaired for nodulation with soybean and Cajanus cajan. Mol**. Plant Microbe Interact. 19:43–52. 2006.

PIRO, V. C., FAORO, H., WEISS, V. A., STEFFENS, M. B. R., PEDROSA, F. O., SOUZA, E. M., e RAITTZ, R. T. **"Open Access FGAP : an automated gap closing tool,"** pp. 1–5, 2014.

QUISPEL, A. **A search of signals in endophytic microor-ganisms**. In: Verma, D.P.S. (Ed.), Molecular Signals in Plant–Microbe Communications. CRC Press, Boca Raton, FL, pp. 471–491. 1992.

RABUS, R., KUBE, M., HEIDER, J., BECK, A., HEITMANN, K., WIDDEL, F., e REINHARDT, R. **"The genome sequence of an anaerobic aromatic-degrading**

denitrifying bacterium, strain EbN1.," Arch. Microbiol., vol. 183, no. 1, pp. 27–36, Jan. 2005.

REINHOLD-HUREK, B. e HUREK, T. **The Genera Azoarcus, Azovibrio, Azospira and Azonexus. In The Prokaryotes**: A Handbook on the Biology of Bacteria, 3rd edn, vol 5, pp. 873-891. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer e E. Stackebrandt. New York, NY: Springer. 2006.

REINHOLD-HUREK, B., HUREK, T., GILLIS, M., HOSTE, B., VANCANNEYT, M., KERSTERS, K. e DE LEY, J. **Azoarcus gen. nov., nitrogenfixing Proteobacteria associated with roots of Kallar Grass (Leptochloa fusca L. Kunth), and description of two species, Azoarcus indigens sp. nov. and Azoarcus communis sp. nov**. Int J Syst Bacteriol 43, 574-584. 1993.

ROY, A. S., BARUAH, R., GOGOI, D., BORAH, M., SINGH, A. K., DEKA, B. H. P. **Draf genome sequence of Pseudomonas aeruginosa strain N002, isolate from crude oil-contaminated soil from Geleky, Assam, India**. Genome Annouc, 1(1):e00104-12, 2013.

SEEFELDT, L. C., HOFFMAN, B. M., e DEAN, D. R. **Mechanism of Mo-dependent nitrogenase. Annual Review of Biochemistry**, 78, 701–22. doi:10.1146/annurev.biochem.78.070907.103812. 2009.

SONG, B. e HÄGGBLOM, M. **"Taxonomic characterization of denitrifying bacteria that degrade aromatic compounds and description of Azoarcus toluvorans sp. nov. and Azoarcus toluclasticus sp.,"** Int. J. …, no. 1 999, 1999.

SONG, B., HÄGGBLOM, M. M., ZHOU, J., TIEDJE, J. M. e PALLERONI, N. J. **Taxonomic characterization of denitrifying bacteria that degrade aromatic compounds and description of Azoarcus toluvorans sp. nov. and Azoarcus toluclasticus sp. nov**. Int. J. Syst. Bacteriol 49, 1129-1140. 1999.

SPRINGER, N., LUDWIG, W., PHILLIP, B. e SCHINK, B. **Azoarcus anaerobius sp. nov., a resorcinol-degrading, strictly anaerobic, denitrifying bacterium**. Int. J.. Syst. Bacteriol 48, 953-956. 1998.

TAGO, K., ISHII, S., NISHIZAWA, T., OTSUKA, S., SENOO, K. **Phylogenetic and functional diversity of denitrifying bacteria isolated from various rice paddy and rice-soybean rotation fields**. Microbes Environ. 26:30–35. 2011.

VIALLE, R. A. S., **"Um sistema para anotação automática de genomas utilizando técnicas independentes de alinhamento.,"** Set. Educ. Prof. e Tecnológica, Univ. Fed. do Paraná, Curitiba. 2013., 2013.

ZERBINO, D., e BIRNEY, E. **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. Genome Research, v. 18, p. 821-829, 2008.

ZHOU, J., FRIES, M. R., CHEE-SANFORD, J. C. e TIEDJE, J. M. **Phylogenetic analyses of a new group of denitrifiers capable of anaerobic growth on toluenevand description of Azoarcus tolulyticus sp. nov**. Int. J. Syst. Bacteriol 45, 500-506. 1995.

KRAUSE, A., RAMAKUMAR, A., BARTELS, D., BATTISTONI, F., BEKEL, T., BOCH, J., GOESMANN, A. **Complete genome of the mutualistic, N2-fixing grass endophyte Azoarcus sp. strain BH72.** Nature Biotechnology, 24(11), 1385–91. doi:10.1038/nbt1243. 2006.

APÊNDICES

1. SCRIPT DE QUEBRA EM 500 PB

O “Script de quebra em 500 pb”, foi desenvolvido pelo doutorando do programa de ciências biológicas da UFPR, Rodrigo Cardoso (não publicado). Como o programa de montagem Newbler trabalha com leituras curtas, de no máximo 500 pb, foi necessário o uso deste script para realizar fragmentação de *contigs* longos gerados através de montagens anteriores. Esse script fragmenta os *contigs* em tamanho de 500 pb, mas mantendo uma sobreposição de 50 bases para permitir a remontagem dos *contigs* originais. O script foi desenvolvido na linguagem de programação *Phyton*.

```
input = open("arquivo_fasta_dos_reads.fasta", "r")
output = open("arquivo_fasata_de_saida.fasta", "w")
line_ant = "AAAAAA"
cab = ">AXYZ"
n = 0
for line in input.readlines():
    if line[0] == ">":
        print n
        if len(line_ant) > 500:
            i = 0
            while i <= len(line_ant)-500:
                output.write(cab[:-1] + "_" + str(i))
                output.write("\n")
                output.write(line_ant[i:i+450])
                output.write("\n")
                i = i + 50
            cab = line.split("  ")[0]
            line_ant = ""
            i = 0
```

```
        else:
            output.write(cab)
            #output.write("\n")
            output.write(line_ant)
            output.write("\n")
            cab = line
            line_ant = ""
        n = n + 1
    else:
                line_ant = line_ant + line[:-1]
```

2. COMANDOS UTILIZADOS NO VIM PARA FORMATAÇÃO DE CABEÇALHOS

Para substituir " 1:N:0:3" por "_1:N:0:3" e " 2:N:0:3" por "_2:N:0:3".

Sintaxe de utilização do no vim:
:%s/PALAVRA_ORIGINAL/PALAVRA_NOVA/g

Sintaxe que representa um espaçamento com uma *string*:
_s\+*string*

Exemplo utilizado:
_s\+1:N:0:3

Comando final:
:%s/_s\+1:N:0:3/_1:N:0:3/g

3. SCRIPT PRÉ-NEWBLER

Foi necessário o uso deste script para realizar alterações no formato dos arquivos fasta que estavam no formato *new-style*. Este formato o newbler não lê, o script faz a mudança para o modelo antigo. O script segue abaixo:

```
cat arquivo_de_reads.fasta | awk '{if (NR % 4 == 1) {split($1, arr, ":"); printf "%s_%s:%s:%s:%s:%s#0/%s (%s)\n", arr[1], arr[3], arr[4], arr[5], arr[6], arr[7], substr($2, 1, 1), $0} else if (NR % 4 == 3){print "+"} else {print $0} }' > arquivo_desaida.fasta
```

4. SCRIPT GERA DATASET DE RETIRADA DE “N”

O “Script gera *DATASET* de retirada de N”, desenvolvido pelo próprio autor, foi utilizado para remover as bases ambíguas do conjunto de dados submetido ao FGAP (PIRO et al, 2014). As montagens não utilizadas na montagem principal foram reunidas e realizado a retirada das bases ambíguas, bases “N”, nas regiões de *gaps*. Estas regiões foram primeiramente agrupadas em um único N e posteriormente removidos, repartindo o *contig* em um novo *contig*. O script foi desenvolvido em Matlab da *The mathworks*.

```
function ioquebraN_RRM(filename)

%Para carregar a sequencia, inserir o nome do arquivo ou caminho entre
%aspas simples '

seq = fastaread(filename);
seqN = quebractsyRRM(seq);

tmp = cellfun(@(x) strrep(x,'N',''), {seqN.Sequence}, 'UniformOutput', false);
for i = 1:length(tmp)
   seqN(i).Sequence = tmp{i};
end
```

```
%Pode ser determinado o numero minimo de bases que possam ser considerados
%uma sequencia, no caso esta 5 como padrao
index = cellfun(@(x) isempty(x) | length(x)<5, {seqN.Sequence});
seqsv = seqN(~index)

filenew = strrep(filename, '.fa', '');
fastawrite([filenew '.semN.fa'],seqsv);
clear
```